Water Resources Research

RESEARCH ARTICLE
10.1029/2020WR027692

Toward Global Stochastic River Flood Modeling

Oliver E. J. Wing[1,2], Niall Quinn[2], Paul D. Bates[1,2], Jeffrey C. Neal[1,2], Andrew M. Smith[2], Christopher C. Sampson[2], Gemma Coxon[1], Dai Yamazaki[3], Edwin H. Sutanudjaja[4], and Lorenzo Alfieri[5,6]

[1]School of Geographical Sciences, University of Bristol, Bristol, UK, [2]Fathom, Bristol, UK, [3]Institute for Industrial Science, University of Tokyo, Tokyo, Japan, [4]Department of Physical Geography, University of Utrecht, Utrecht, Netherlands, [5]CIMA Research Foundation, Savona, Italy, [6]European Commission Joint Research Centre, Ispra, Italy

Abstract Global flood models integrate flood maps of constant probability in space, ignoring the correlation between sites and thus potentially misestimating the risk posed by extreme events. Stochastic flood models alleviate this issue through the simulation of flood events with a realistic spatial structure, yet their proliferation at large scales has historically been inhibited by data quality and computer availability. In this paper, we show, for the first time, the efficacy of modeled river discharge reanalyses in the characterization of flood spatial dependence in the absence of a dense stream gauge network. While global hydrological models may show poor correspondence with absolute observed river flows, we find that the rate at which they can simulate the joint occurrence of relative flow exceedances at two given locations is broadly similar to when a gauge-based statistical model is used. Evidenced over the United States, flood events simulated using observed gauge data from the U.S. Geological Survey versus those generated using modeled streamflows have similar (i) distributions of site-to-site correlation strength, (ii) relationships between event size and return period, and, importantly, (iii) loss distributions when incorporated into a continental-scale flood risk model. Extremal dependence is generally quantified less accurately on larger rivers, in arid climates, in mountainous terrain, and for the rarest high-magnitude events. However, local-scale errors are shown to broadly cancel each other out when combined, producing an unbiased flood spatial dependence model. These findings suggest that building accurate stochastic flood models worldwide may no longer be a distant aspiration.

Plain Language Summary Global flood risk is commonly estimated through flood inundation maps with a defined probability of occurrence. These flood simulations have a key drawback in that they fail to capture the spatial patterns exhibited during real flood events, instead modeling the same probability of flooding on every river at once. Solutions which rely on networks of gauged river flow observations will necessarily break down in the majority of the world's regions which lack such a resource. In this paper, we use historic river flows simulated by global rainfall-runoff models (rather than observed flows) into a statistical model which captures the spatial correlation of flow extremes. If we examine the relative flow exceedance probabilities from these hydrological models rather than the volumetric flow values, flood events are generated which exhibit similar characteristics to those when gauged flow observations are used. Crucially, the simulation- and observation-generated flood events produce near-identical losses to buildings in the United States. The implications of this are that true stochastic flood risk models, which account for spatial dependence, can proliferate globally via the generation of realistic flood event sets from hydrological models.

Key Points:
- Large-scale flood hazard models typically neglect to represent the spatial dependence of real flood events
- Relative flow exceedances simulated by the fusion of global hydrological with statistical models reproduce gauge-driven flood event sets
- At the continental scale, key characteristics of a flood risk model are indistinguishable when driven with observed versus modeled flow data

Supporting Information:
- Supporting Information S1

Correspondence to:
O. E. J. Wing,
oliver.wing@bristol.ac.uk

Citation:
Wing, O. E. J., Quinn, N., Bates, P. D., Neal, J. C., Smith, A. M., Sampson, C. C., et al. (2020). Toward global stochastic river flood modeling. *Water Resources Research*, *56*, e2020WR027692. https://doi.org/10.1029/2020WR027692

Received 9 APR 2020
Accepted 15 JUL 2020
Accepted article online 19 JUL 2020

1. Introduction

Hydraulic modeling at large spatial scales is a field of enquiry approaching a state of maturity, with the flood maps produced beginning to inform wide-area planning decisions, insurance pricing, and emergency response. Historically a reach-scale venture, multiple institutions from academia, industry, and elsewhere have expanded the spatial domain they consider when developing models of fluvial flood inundation up to the entire globe, as a response to the scarcity of flood hazard data in most world regions (Dottori et al., 2016; Pappenberger et al., 2012; Sampson et al., 2015; Winsemius et al., 2013; Yamazaki et al., 2011). These flood maps, however, are typically "static"; that is, they are a spatially homogeneous representation of a given probability flood. While at the local scale this may be representative of a realistic

event (e.g., a small stretch of river experiencing a *T*-year flood event), as the size of the area considered increases, the resultant flood extents become increasingly unrealistic when viewed as a simultaneous phenomenon (e.g., an entire country experiencing a *T*-year flood event at the same time). Actual floods vary in their extremity across space: If a given location is extreme, you may expect proximal locations to be similarly extreme and distal locations to be decreasingly extreme (Keef et al., 2013). This asymptotic spatial dependence structure reflects the spatial heterogeneity of the driving physical processes in the atmosphere and how the terrestrial (sub)surface responds hydrologically (Blöschl & Sivapalan, 1997; Smith et al., 1994). Univariate (at-a-point) flood frequency analysis does not capture this spatial structure. Yet, current national-scale risk mapping is generated through assimilating a mosaic of thousands of local-scale univariate models (Gilles et al., 2012), for example, by the U.S. Federal Emergency Management Agency (FEMA) and the England Environment Agency. Equally, the simultaneous simulation across space of univariately defined probability floods is how typical large-scale flood hazard models are executed (Alfieri et al., 2014; Dottori et al., 2016; Pappenberger et al., 2012; Sampson et al., 2015; Wing et al., 2017; Winsemius et al., 2013; Yamazaki et al., 2011).

The failure to represent spatial dependence in risk calculations often leads to a misestimation of wide-area risk profiles (the probability-loss curve) when quantified using such data (Alfieri et al., 2016; Hirabayashi et al., 2013; Ward et al., 2013; Wing et al., 2018; Winsemius et al., 2016). While spatially constant probability flood maps can be used to estimate average annual losses (by integrating the area under the probability-loss curve), they cannot tell us about losses which may occur in more extreme loss years. Such events are the very thing insurers must quantify for pricing (and to ensure they have access to adequate capital for regulatory and solvency reasons), that emergency managers must prepare for, and that corporations require to mitigate supply-chain risk. Metin et al. (2020), using a flood model with realistic spatial dependence for the Elbe basin, Germany, demonstrate that damages are underestimated by static flood maps for return periods more frequent than 1 in 50 years and overestimated for return periods less frequent than 1 in 200 years. They found the tail of the loss distribution (>200-year) was at least doubled when conflating the *T*-year discharge with the *T*-year loss. This erroneous conflation and its consequences have been further documented by Lamb et al. (2010), Wyncoll and Gouldby (2015), and Thieken et al. (2015).

Such findings support a move away from the spatially homogeneous risk maps that are ubiquitous in flood risk management toward an approach that accounts for the spatial structure of real flood events (Vorogushyn et al., 2018). Although understudied relative to the engineering-standard univariate case, multivariate flood frequency analyses have proliferated since the turn of the century. The characterization of spatio-temporal extreme flows is commonly performed either by continuous hydrological simulation (Falter et al., 2015, 2016; Grimaldi et al., 2013; Haberlandt & Radtke, 2014) or fitting statistical dependence models to samples of stream gauges (Diederen et al., 2019; Ghizzoni et al., 2012; Keef et al., 2009, 2013; Neal et al., 2013; Quinn et al., 2019). The former involves cascading synthetic meteorological time series through a rainfall-runoff model, from which the output synthetic discharge time series can form the basis of a flood frequency analysis. This approach has the advantage of full coverage of the river network (at least, down to the resolution of the model) and simulating the full event hydrograph in preparation for unsteady hydraulic modeling, while an observation-based statistical model contends with gauge density issues and simulates only the joint occurrence of flood peaks. However, continuously simulating enough years (10,000 years is common) to adequately explore tail flood risk comes at high computational cost: lower process representation, coarser grid scales, and smaller model domains are necessitated across the model cascade. Furthermore, the use of hydrologically modeled discharges to drive hydraulic inundation models at large spatial scales is subject to substantial uncertainty (Beck et al., 2017; Prudhomme et al., 2011; Sperna Weiland et al., 2010). Such flows are often generated amidst a lack of data with which to calibrate or parameterize the hydrological models (Beven & Cloke, 2012) and are driven with meteorological data that may poorly represent extremes (Beck et al., 2019; Kendon et al., 2017; Kidd et al., 2012). These uncertainties can be exacerbated by the hydraulic modeling component (Falter et al., 2016; Grimaldi et al., 2019), cascading into loss estimates that are thus highly sensitive to the simulated flood peaks (Sampson et al., 2014; Zischg et al., 2018). For the purposes of spatial dependence modeling, not only is the accuracy of simulated discharge at-a-point of concern, so too is the ability of the hydrological model to accurately simulate the occurrence of concomitant flood peaks elsewhere. Continuous hydrological simulation studies to date are yet to evidence the validity of the joint occurrence and magnitude of modeled extreme discharges at-scale.

Adopting a gauge-based multivariate statistical modeling approach reduces accuracy issues through the use of observations of river discharge, though are still subject to uncertainties arising from measurement error, the stage-discharge relationship, and the extrapolation of short time series (McMillan et al., 2012). Such an approach also minimizes computational load through avoiding rainfall-runoff modeling and simulating the hydraulic extremes only. This permits hydraulic modeling of greater complexity and higher resolution applied across wider regions. The perennial issue in hydrology still remains, however: How do you characterize flows on ungauged streams? For the univariate case, flood frequency analyses have been regionalized by correlating extreme flow observations with catchment characteristics, and assuming unobserved flood behavior in similar catchments is thus predictable from the regionalization (Merz & Blöschl, 2009; Salinas et al., 2013; Smith et al., 2015). This time-invariant approach is, however, not applicable for modeling spatial dependence. Quinn et al. (2019), in the only continental-scale stochastic flood risk model subjected to peer-review (models similar in concept exist in the insurance industry but remain scientifically opaque), adopt a crude interpolation method whereby single river gauges transpire to represent large ungauged areas. While this may be acceptable in the gauge-rich contiguous United States (CONUS), reliably applying the Quinn et al. (2019) gauged-based model across a mostly data-poor planet is not possible (Hannah et al., 2011). Ultimately, to obtain a view of flood spatial dependence at the global scale, simulated discharge from large-scale hydrological models will necessarily play a role.

In this paper, we examine the extent to which simulated discharge time series from global hydrological models can reliably inform a stochastic view of flood risk at large spatial scales. To do so, we emulate the gauge-based approach of Quinn et al. (2019) in the CONUS, its statistical engine underpinned by the conditional exceedance model of Heffernan & Tawn (2004; hereafter, H&T), but replace the observational flow inputs with those simulated by global hydrological models. We compare the characteristics of the model-driven H&T synthetic events with those generated by a gauge-driven H&T model in terms of their spatial patterning and degree of extremity. We then produce event depth grids by linking the synthetic return periods at each gauge location for each event and the presimulated suite of CONUS-wide return period flood maps presented in Wing et al. (2017). In this way, the often inaccurate absolute discharges from global hydrological models are disregarded in favor of their exceedance probabilities, while the channel hydraulics and flood inundation dynamics are still informed by a gauge-based flow estimation procedure. With an inundation map for every event, we can examine the risk profile of a sample building inventory when run against the observation- and simulation-driven models and compare their similarity in the context of vulnerability estimation uncertainty. This analysis thus demonstrates, for the first time, the efficacy of simulated streamflow in large-scale stochastic modeling, illustrating the extent to which scientifically sound studies of flood spatial dependence globally may proliferate.

2. Methods

The core concepts underpinning the research presented in this paper are illustrated in Figures 1 and 2, providing a small-scale example of the kinds of analyses we are subsequently undertaking at the continental scale in section 3. These figures show an exemplar analysis for two sites in the U.S. Midwest—Platte River at Louisville, NE, and, 250 km downstream, Missouri River at St. Joseph, MO (see Figure 1a)—where information on the magnitude and frequency of flows in time and space from a global hydrological model and river gauges are contrasted. Full methodological details are presented in subsequent subsections. Figures 1b and 1c depict a model hindcast and historic observations of flow at the two sites, respectively. Noting the y axis scale in Figure 1b is double that of Figure 1c, it is evident that the hydrological model severely overestimates discharge at each location. The 1 in 5-year discharge (20% annual chance of being exceeded), as computed by fitting a generalized extreme value (GEV) distribution to annual maxima, on the Platte River is 1,800 m^3s^{-1} based on observations compared to 4,300 m^3s^{-1} according to the model. Similarly, the 5-year discharge on the Missouri River is 4,900 m^3s^{-1} when calculated using gauged flows, compared to 12,200 m^3s^{-1} when using the model. Equally, there is often model-observation disagreement in the chronology of local 5-year flow exceedance. Observed exceedances on the Platte River in 1983 and 1987 according to the river gauges were not simulated by the models, while unobserved-yet-modeled exceedances occurred in 1985, 1998, 2007, and 2014. However, these above findings—which, by most conventional hydrological metrics, would be heavily penalized in any evaluation of model performance—may not be relevant in a characterization of the correlation between the Platte River and the Missouri River:

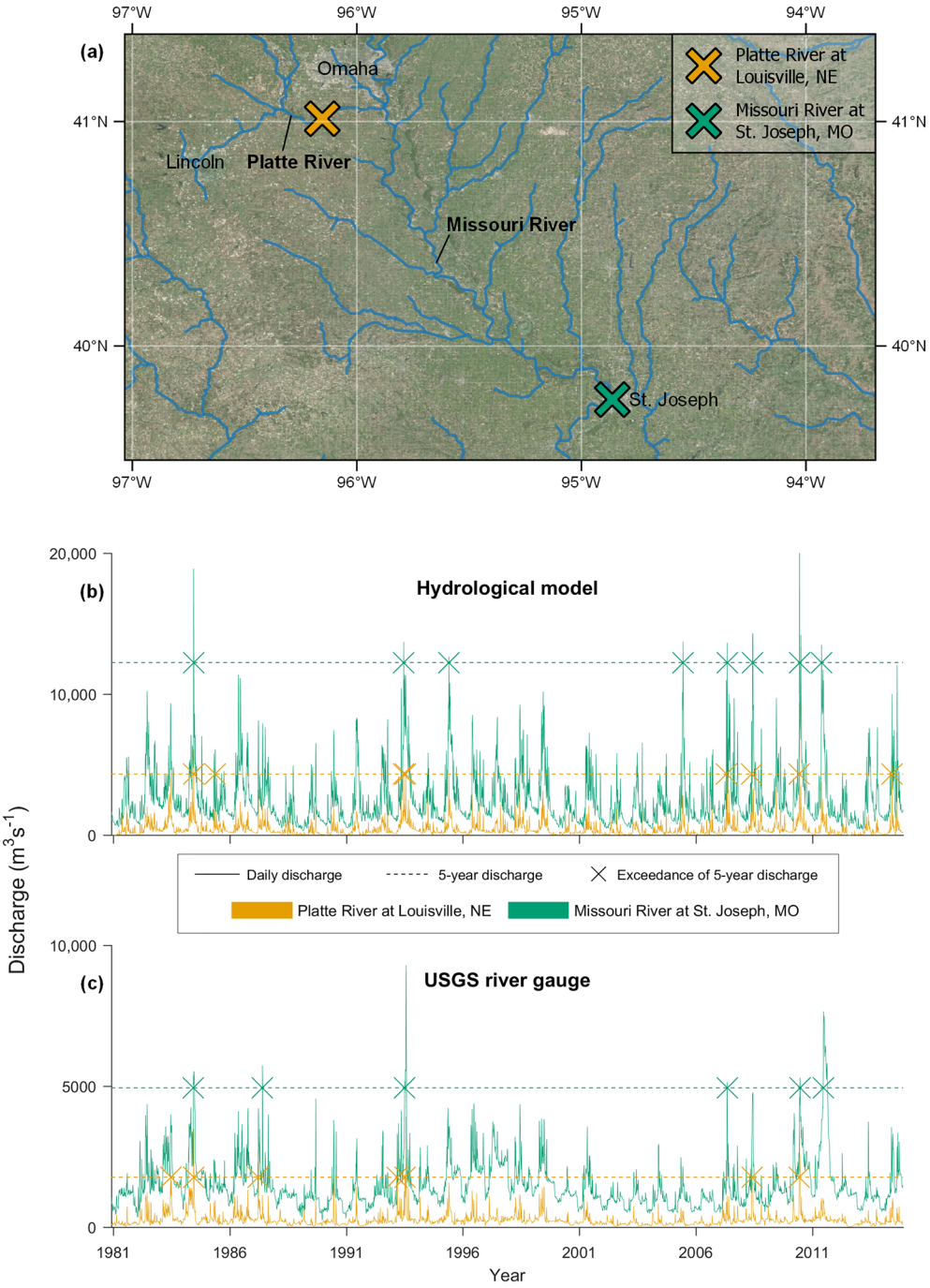

Figure 1. Example analysis comparing global hydrological model output to gauged observations of river discharge at a conditioning site on the Platte River at Louisville, NE, and its neighboring site 250 km downstream on the Missouri River at St. Joseph, MO. Panel (a) shows the locations of the gauging stations; rivers with a drainage area >500 km^2 are shown in blue. Panel (b) shows the PCR-GLOBWB hindcast simulation, and panel (c) shows observations from USGS river gauges of historical flows at both sites, with instances of exceedance of the local, source-specific 5-year discharge highlighted.

that is, the tendency for both sites to simultaneously be extreme. The Platte River exceeds its data source-specific 5-year flow in eight and seven discrete events according to the model and observations, respectively. The Missouri River concurrently exceeds its data source-specific 5-year flow during four of the eight events in the model record and four of the seven events in the gauge record. Hence, at the

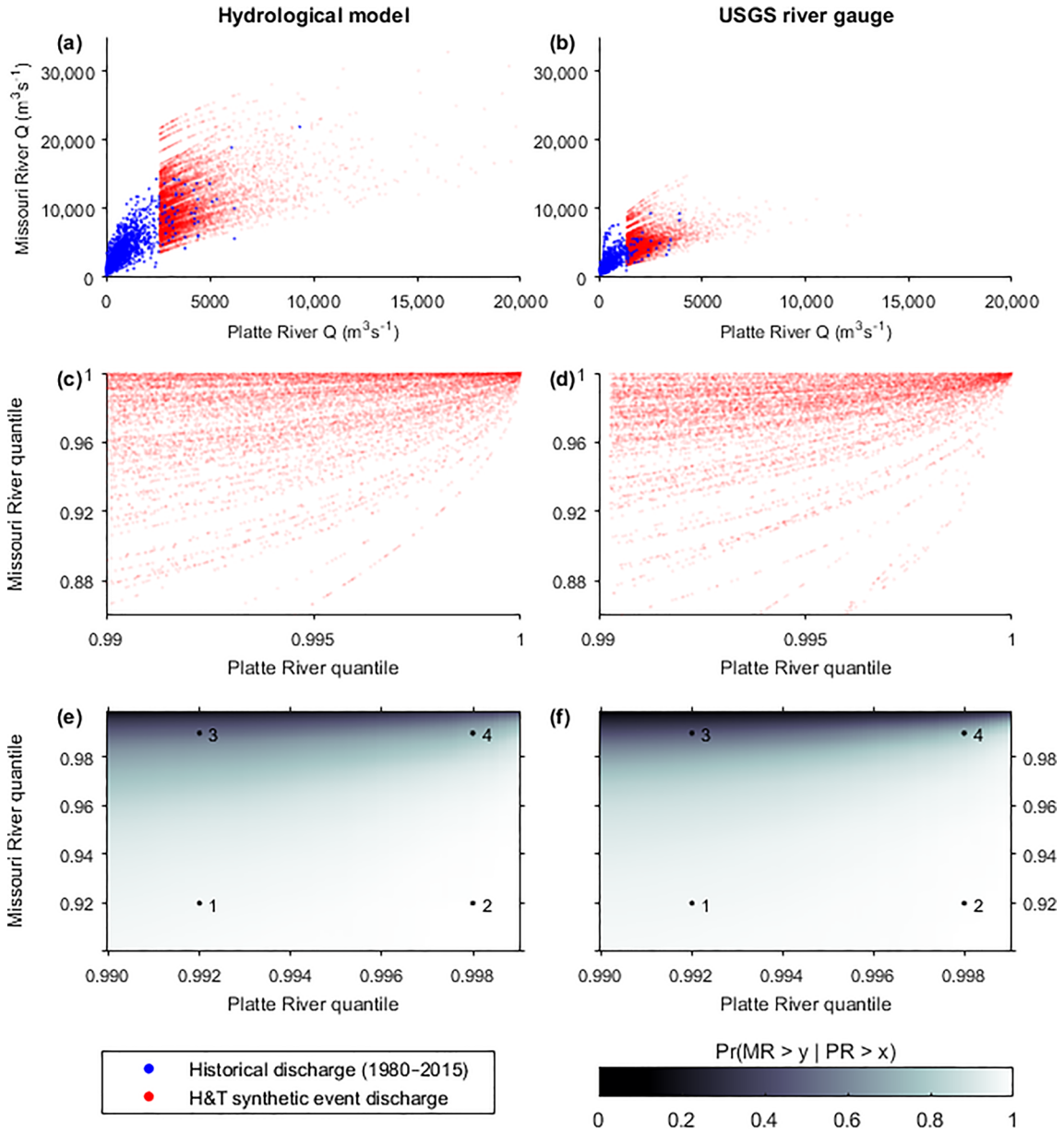

Figure 2. Statistical modeling of modeled and observed discharges on the Platte River at Louisville, NE, and the Missouri River at St. Joseph, MO. Panels (a) and (b) show historic concurrent maximum 9-day discharges (Q) at the conditioning (*x* axis) and neighboring (*y* axis) sites (in blue), as well as 10,000 samples of the fitted H&T joint distribution above a given extreme threshold (96th quantile) at the conditioning site (in red) for the hydrological model and observations, respectively. Panels (c) and (d) show the same synthetic events in panels (a) and (b) but converted to the quantiles of their respective semiempirical flow distributions for the hydrological model and observations, respectively. Panels (e) and (f) are a version of panels (c) and (d) where the probability of the Missouri River (MR) exceeding its respective flow quantile *y* given that the Platte River (PR) exceeds its respective flow quantile *x* is plotted as a surface for the hydrological model and observations, respectively. The probabilities at the numbered points are shown in Table S1 in the supporting information.

arbitrary 5-year frequency threshold, the sites are 57% correlated according to the model and 50% correlated according to the gauges: There is a 57% (model) or 50% (gauge) chance of the Missouri River exceeding its 5-year discharge given that the Platte River has also exceeded its 5-year discharge. In *this* context, the model and the observations have produced very similar results.

We can explore this extreme flow dependency more fully by computing the magnitude and strength of the correlation between the Platte River and the Missouri River using the H&T statistical model (see section 2.2 for more details), from which we can sample any number of plausible flow co-occurrences (here denoted as an "event"). Figure 2a shows, in blue, the same data as in Figure 1b but with concurrent flows at

the two sites plotted against one another. Figure 2b shows the same, but for the gauged discharge observations (note the common x and y axes in Figures 2a and 2b, illustrating the gulf in flow magnitude between model and observations). The red dots depict 10,000 synthetic flood events as simulated by the H&T model, when supplied with modeled (a) and observed (b) flows. Since we are only interested in the correlation of extreme flows, we condition the H&T model on the Platte River exceeding an arbitrary extreme threshold (96th flow quantile). Hence, given Platte River is extreme, it is often the case that the Missouri River 250 km away is also extreme at some time later during the same event—and as Platte River gets more extreme, so the likelihood the Missouri River is also extreme increases. Both Figures 2a and 2b agree on this assertion, though for very different magnitudes of discharge.

Figures 2c and 2d plot those same synthetic flows (red dots in Figures 2a and 2b) but as their respective quantiles on the source- (model/gauge) and site- (Platte River/Missouri River) specific marginal, semiempirical distribution, or, equally, the cumulative probability of flow exceedance. Here, we can see that the model- (c) and gauge- (d) based simulated flow quantiles are broadly similar: Generally, the Missouri River is often extreme when the Platte River is too. We can examine this in more detail in the surface plots of Figures 2e and 2f. These are based on the same data as Figures 2c and 2d but illustrate the probability that the Missouri River exceeds the quantile on the y axis given that the Platte River exceeds the quantile on the x axis. The probabilities at the numbered points are shown in Table S1. Again, we see very similar results for both input data sources in terms of the probability that the two sites experience certain flow conditions.

The approach of the work in this paper broadly follows the concepts outlined in Figures 1 and 2 but for all available U.S. Geological Survey (USGS) river gauges (subject to quality control; see section 2.1.1), the hypothesis being that models which substantially misestimate streamflow in absolute terms may still have skill in the characterization of site-to-site correlation in terms of relative exceedance probability; in essence, "normalized" modeled streamflows (expressed as a probability) represent flood spatial dependence similarly to observations. Details on the river gauges, hydrological models, statistical model, and flood risk model are outlined in the following subsections, with tests for the similarity of flood spatial dependence modeling between gauge and hydrological model inputs described. The key steps of this research are outlined in the flowchart of Figure 3.

2.1. Input Data

2.1.1. USGS River Gauges

As per Quinn et al. (2019), we obtained daily river flow time series at approximately 9,800 gauging locations from the USGS. These records are often discontinuous and of variable quality and so are filtered whereby gauges are selected if they (i) contain data for at least 35 years between 1980 and 2015; (ii) have less than 2% erroneous or missing measurement days; and (iii) exhibit no discernible trend or step changes (as determined by the Kendall τ test). This resulted in 3,094 river gauges for use in the proceeding analyses.

2.1.2. Global Hydrological Models

We employ three global hydrological models in this analysis. Daily streamflow hindcasts simulated by each model for the period 1980–2015 were extracted at each gauge location in order to facilitate the model-observation dependence comparison. To ensure the hydrological model output is analogous to each respective gauge (grid coarseness and/or errors in the underlying hydrography data may produce a divergence in the size of the simulated and observed river), the drainage area upstream of each gauge (as reported by the USGS) needed to be within 30% of that in the model hydrography. To ensure a fair intercomparison, gauge locations were only retained where this condition was met for all three models: of the 3,094, 1,674 gauges were retained. Unsurprisingly, the rejected gauges were generally on smaller rivers (median drainage area of rejected gauges = 450 km^2), meaning the median drainage area of 1,300 km^2 from the original 3,094 sites increased to 3,400 km^2 in the accepted set of 1,674. This focus on larger rivers—necessitated by the fidelity of available wide-area elevation and hydrography data used by global hydrological models—is consistent with the use of daily streamflow observations in this analysis, which may fail to capture the flashier rainfall-runoff response in smaller catchments. Further, the application of large-scale hydrological models in the context of flood modeling is commonly restricted to larger rivers, owing to their coarse grid resolution (Dottori et al., 2016; Pappenberger et al., 2012; Ward et al., 2013; Winsemius et al., 2013; Wu et al., 2014; Yamazaki et al., 2011). The method in this paper can therefore be viewed as an analysis of fluvial flood spatial dependence, while flooding on smaller rivers (and hydrologically isolated local depressions)—better

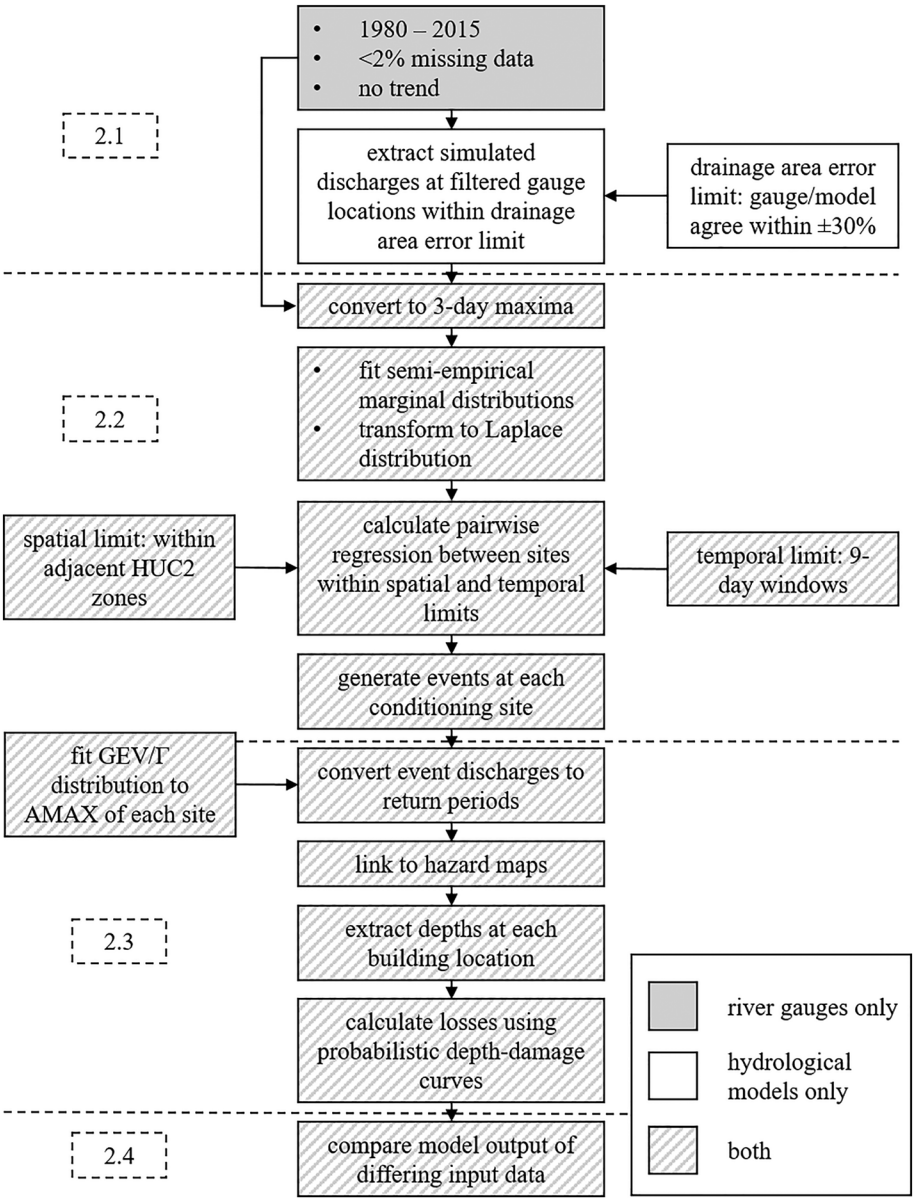

Figure 3. Flowchart of key methodological steps in the building and testing of a stochastic flood model. Dashed boxes contain the subsections relevant to the particular steps.

characterized as pluvial flood events—remain outside the scope of this work. A brief description of the three hydrological models follows.

2.1.2.1. CaMa-Flood (UTokyo)

CaMa-Flood (Catchment-based Macro-scale Floodplain model) is a hydrodynamic global river routing model, which was developed mainly as a river module of global climate models and also as a tool for large-scale flood risk assessment (Yamazaki et al., 2011, 2012). It represents floodplain inundation dynamics as a subgrid physical process based on high-resolution topography data and calculates river discharge using a simplified shallow water flow equation (Yamazaki et al., 2013) along a coarse-resolution river network map including river channel bifurcation (Yamazaki et al., 2014). For this study, we used the latest version of CaMa-Flood (v3.95), which is based on the topography and hydrography data sets MERIT DEM and MERIT Hydro (Yamazaki et al., 2017, 2019). The model was run, uncalibrated, at 6 arc minutes (~11 km at the equator) spatial resolution from 1980–2015 forced by the daily runoff product calculated by the HTESSEL land surface model for the Earth2Observe project (Schellekens et al., 2017).

2.1.2.2. PCR-GLOBWB (UUtrecht)

PCR-GLOBWB (PCRaster Global Water Balance; van Beek et al., 2011; Sutanudjaja et al., 2018) is an open-source global hydrology and water resource model that has been used for various water-related global change studies, such as the impact of land use change on global water resources (Bosmans et al., 2017), water scarcity and drought (Wanders & Wada, 2015), groundwater depletion (de Graff et al., 2019), as well as for a myriad of flood-related studies. Examples include medium-range to seasonal flood forecasting (Candogan Yossef et al., 2013), future flood events (Sperna Weiland et al., 2012), and current and future flood hazard and risk (Ward et al., 2013; Winsemius et al., 2016). For this study, we implemented the latest version, PCR-GLOBWB 2 (Sutanudjaja et al., 2018), at the spatial resolution of 5 arc minutes (~10 km at the equator). The meteorological forcing files were based on monthly CRU TS 3.2 (Harris et al., 2014), daily ERA-40 (Uppala et al., 2005), and ERA-Interim (Dee et al., 2011) for hindcast simulations of the period 1958–2015. No calibration was performed to the standard parameterization sets. The simulation setup for this study used an advanced surface water kinematic wave routing scheme, which includes floodplain inundation (see, e.g., Winsemius et al., 2013).

2.1.2.3. GloFAS (JRC/ECMWF)

The third considered data set is the v3.0 reanalysis of the Global Flood Awareness System (GloFAS; Alfieri et al., 2020). It includes estimates of daily river discharge for 1980–2018 with quasi-global coverage (90°N to 60°S) and a spatial resolution of 6 arc minutes (~11 km at the equator). The GloFAS streamflow reanalysis was produced with the LISFLOOD hydrological model (van der Knijff et al., 2010) and ECMWF's atmospheric reanalysis ERA5 (Herzbach et al., 2018) as meteorological input. In the GloFAS-Reanalysis v3.0, LISFLOOD was calibrated using at least 4 years of observed discharges at 1,226 stations in 66 countries, of which 278 were in the United States. GloFAS is designed for medium to large river basins with drainage area larger than 5,000 km^2. Performance at the 278 calibration stations in the United States have median scores of 0.66 Kling-Gupta Efficiency, 0.75 correlation, and 5.5% bias (Alfieri et al., 2020).

2.2. H&T Model

In order to define independent events, we need to be explicit in our meaning of "simultaneous," since different sites may not necessarily experience peak flows from the same event on the same day. Following Quinn et al. (2019), we consider a temporal window of 9 days: suitably long to capture likely flood wave travel times during a single event, and suitably short so as to avoid capturing spurious correlations. However, this time window would require a given conditioning site to be regressed against every gauge nine times (once for each day in the window), presenting a cumbersome computational problem. As such, we temporally coarsen the input data to 3-day maxima so that these time-lagged regressions are required only three times for each neighboring site. To further reduce the number of required joint regressions and prevent unrealistic, spurious correlations, we apply spatial limits to which gauge locations can be considered dependent. Following Quinn et al. (2019) again, sites may only be correlated within large hydrological regions (defined by Level 2 USGS HUC basins): the region the conditioning sites falls within and those it is adjacent to (see Figure 4a). These arbitrary spatial and temporal limits are flexible, yet the choices made are consistent with our current understanding of the typical duration and distribution of U.S. flood events (e.g., Allen et al., 2018; Smith et al., 2011). Further, Quinn et al. (2019) demonstrate these assumptions produce realistic patterns of flooding across the United States.

The statistical model underpinning the dependence characterization was first presented in Heffernan and Tawn (2004), and its application to stochastic flood modeling is detailed more exhaustively in Keef et al. (2009, 2013), Neal et al. (2013), Wyncoll and Gouldby (2015), and Quinn et al. (2019). First, a marginal distribution is fitted to the flow data at each location, defining the probability a certain discharge is exceeded at the given site—independent of flows occurring at neighbors. These distributions are "semiempirical," whereby the original (empirical) data characterizes exceedance probabilities below a given quantile threshold and a generalized Pareto distribution is fitted to data above this threshold to define the tails. This step requires a careful balance between a quantile threshold low enough to have a suitable number of data points with which to estimate the generalized Pareto model (i.e., to keep variance low) yet high enough to avoid model misspecification bias. This "bias-variance trade-off" is guided by the cross-validatory threshold selection method of Northrop et al. (2017). The modal threshold across all sites and data sources was the 96th

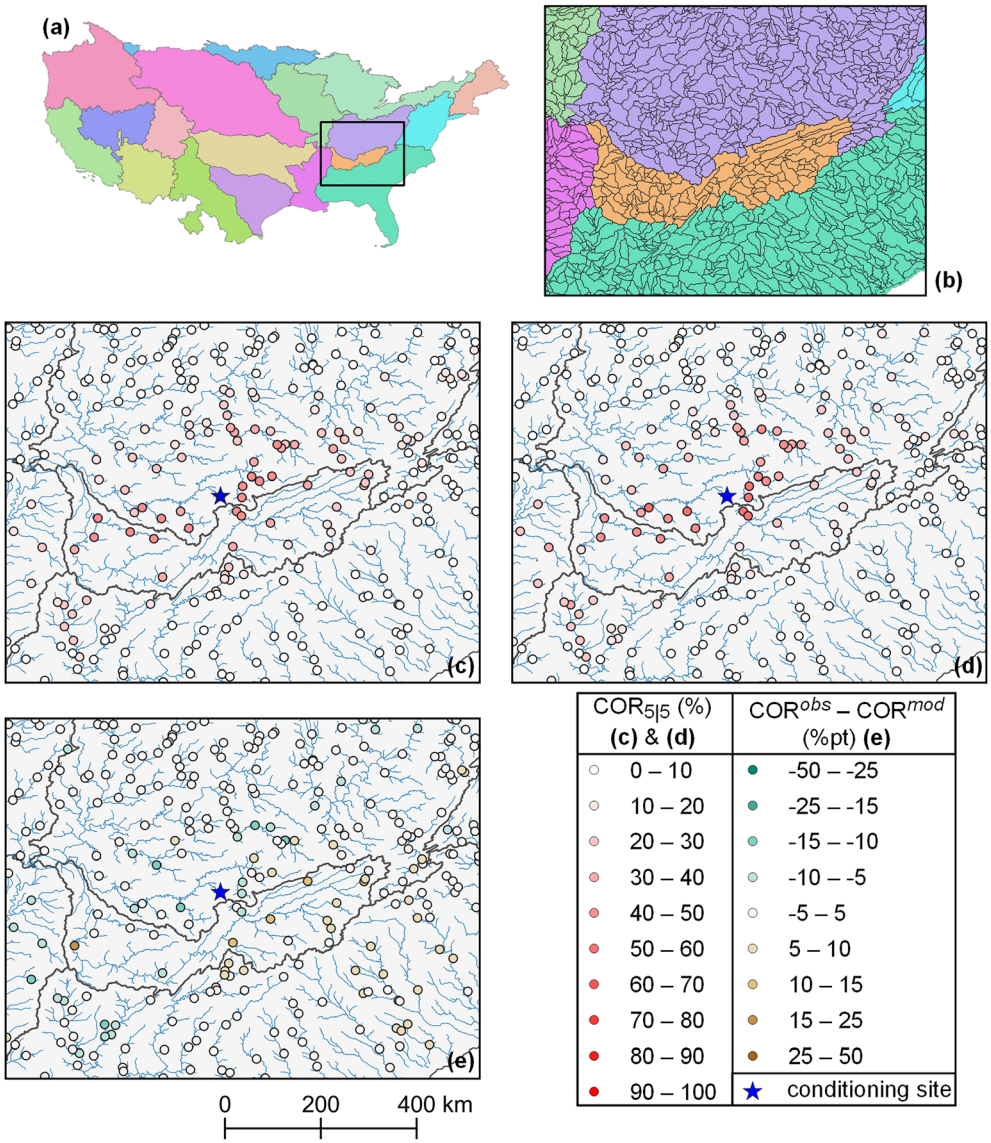

Figure 4. Hydrologic discretization of the United States with H&T model output summarized at an example conditioning site on the East Fork Obey River near Jamestown, TN. Panel (a) shows the 18 HUC2 zones used to define the spatial limits within which an event can occur; panel (b) shows the HydroBASINS (Level 8) used to extract event-specific inundation maps from presimulated hazard data; panels (c) and (d) show event co-occurrence rates (COR; see Equation 2), where an event is defined, in this instance, as the conditioning site (in blue) and a given neighbor (white–red) exceeding the 5-year flow, when simulated using USGS gauge and CaMa-Flood input discharge data, respectively; panel (e) shows the difference between gauge- (^{obs}) and model- (^{mod}) based event co-occurrence rates. Blue lines represent river channels with drainage area exceeding 500 km^2.

quantile. These marginal distributions then need to be transformed onto a common scale, in this case, the Laplace distribution, since this induces a linear relationship between joint extremes.

A generic formulation of the H&T model, simulating discharges at neighbor site Y_j given conditioning site Y_i is extreme, can be described as

$$Y_j|(Y_i = y) = \alpha_{j|i} y + y^{\beta_{j|i}} Z_{j|i} \quad (1)$$

where y is a vector of discharges at site Y_i that exceed a specified extreme flow quantile on the Laplace margin; Y_j can be every neighbor to the conditioning site within the spatial limits; parameter $\alpha_{j|i}$ controls the overall strength of the dependence at site Y_j given Y_i and falls within the limit [−1,1]; parameter $\beta_{j|i}$

controls how different values of y change this dependence and is limited to <1; and $Z_{j|i}$ is a vector of residual terms, which, by virtue of their independence from variable Y_i, permit nonparametric modeling of $Y_j|(Y_i = y)$ beyond the measured (or, in the case of the hydrological models, simulated) range of the input data. Equation 1 is repeated for the temporal lags at Y_j within the 9-day window.

In this analysis, we repeat Equation 1 for all conditioning sites, each of the conditioning site's neighbors, and each neighbor's temporal lag for 10,000 values of y (extreme flow samples of equal occurrence probability). The maximum lagged streamflow from a given conditioning-neighbor-y combination is extracted, meaning every conditioning site has 10,000 (extreme) discharge samples and every neighbor within the spatial limits has 10,000 corresponding discharge samples—each one of the 10,000 groups of discharges is considered to be an "event." With 10,000 events at each conditioning site and 1,674 sites in total, we thus analyze 16,740,000 synthetic flood events across the United States for each input data source.

2.3. Fathom-US CAT Model

The simulated streamflows from each synthetic event are transformed into a return period derived from the historical site-specific discharges from each input data source (observed or simulated). Maximum likelihood estimation determined the closest fit of either a Gamma or GEV distribution to annual maxima of these historical simulated or observed flows. From these distributions, the synthetic event discharges were converted to return periods. These return periods informed an "offline" extraction of a water depths from presimulated catalog of return period flood maps, which were then intersected with exposure and vulnerability information to calculate losses arising from each event. This flood catastrophe (CAT) model, Fathom-US CAT, is outlined in the following subsections. For more detailed information on the components of this model: extreme flow estimation (Smith et al., 2015), inundation modeling (Sampson et al., 2015; Wing et al., 2017), loss calculations (Wing et al., 2018, 2020), and event set generation (Quinn et al., 2019).

2.3.1. Flood Inundation Model

Spatially continuous flood maps of the CONUS were presented and validated in Wing et al. (2017) based on the global flood modeling methodology of Sampson et al. (2015). Each map, simulated at 1 arc second (~30 m at the equator) resolution, represents one of 17 "constant return period in space" flood layers of the CONUS, ranging between 1 in 5 and a 1 in 1,000 years (20% and 0.1% annual exceedance probability). These probabilities are derived from an estimation of extreme flows using a U.S.-only version of the global regionalized flood frequency analysis (RFFA) of Smith et al. (2015). This approach is underpinned by the assumption that extreme flows on ungauged streams are predictable from their catchment characteristics (Meigh et al., 1997; Salinas et al., 2013). The information from gauged streams of similar characteristics to a given ungauged stream—in terms of annual average rainfall, upstream drainage area, dominant climatology, and average slope—is "transferred" to predict extreme flows on the latter. Median errors in the estimation of a 1 in 100-year discharge globally were reportedly 56%, though the uncertainty in the gauged observation of the 100-year flow itself can be at least 40% due to measurement error and rating curve configuration (Coxon et al., 2015; Di Baldassarre & Montanari, 2009; McMillan et al., 2012). Further, in its application to the inundation model, biases in the RFFA are dampened somewhat by ensuring river channels can convey the 1 in 2-year discharge, a standard geomorphological assumption. The computational hydraulic engine is based on LISFLOOD-FP (Bates et al., 2010; de Almeida et al., 2012), which numerically solves a local inertial formulation of the shallow water equations in 2-D over the USGS National Elevation Dataset in this instance. Further details are available from Wing et al. (2017).

Each synthetic event guides an extraction from this inventory of flood hazard layers, depending on the return period of the flow simulated by the H&T model at a given location. As per Quinn et al. (2019), each gauge location is assigned a representative river catchment—a discrete spatial unit within which we can reasonably expect the return period to remain constant during a given event. To define this, we use the HydroBASINS data set (Lehner & Grill, 2013). Each gauge site is assigned a Level 8 basin or, where multiple gauges fall within the assigned catchment, a Level 10 basin (see Figure 4b). The synthetic flood events at specific points in space (as simulated by the H&T models) are thus converted to inundation depth grids which are a composite sample of varying probability, presimulated flood maps. For more information on the catchment-based sampling procedure, see Quinn et al. (2019) and Wing, Sampson, et al. (2019).

These U.S.-wide inundation maps have been extensively validated against engineering-grade, local-scale inundation models (Wing et al., 2017; Wing, Bates, et al., 2019). The validity of these inundation maps,

however, is not necessarily directly relevant to the analysis in this paper. These are simply used as a means of comparing the risk output of a gauge-based versus a model-based flood event generator. We do not "validate" the event-based depth grids; rather, we are interested in the similarity of the resultant outputs given differing input data rather than interrogating the substantive meaning of the outputs per se.

2.3.2. Loss Calculations

Each synthetic event was intersected with a database of U.S. buildings—the FEMA National Structure Inventory—which contains information on the location, value, and occupancy type of over 100M structures. Event depths were sampled at each structure location, and a probabilistic depth-damage function was applied, depending on the building type, to compute a range of losses to each building during each event as a proportion of its value. These functions are based on the analysis of Wing et al. (2020), who used 2M flood insurance claims in the United States to derive a stochastic relationship between depth and damage. For instance, a flood depth of 0.4 m has a 43%, 28%, 14%, and 15% chance of causing 0–10%, 10–30%, 30–50%, and 50–100% relative damage to a one-story residential building, respectively. At 1.8 m, there is a 21%, 20%, 19%, and 40% chance of causing 0–10%, 10–30%, 30–50%, and 50–100% relative damage, respectively. We sample from these loss distributions 100 times, producing a spread of damages arising from each synthetic flood event. The relative differences in loss between events generated by the different sources of flow data are then compared.

2.4. Model Tests

The meaningful results from this analysis will come from the inter-input data comparison of the spatial footprint and relative magnitude of the synthetic events and the damages they cause. This section describes the tests carried out to interrogate these differences between the gauge- and model-driven structures.

At a given conditioning site (Y_i) for a given input data source, an event co-occurrence rate (COR) is computed under a variety of scenarios for each neighboring site (Y_j):

$$\mathrm{COR}_{p_j|p_i} = \frac{\sum\left(Y_j > p_j | Y_i > p_i\right)}{\sum\left(Y_i > p_i\right)} \times 100 \tag{2}$$

where p is a specified return period and the subscripts i and j indicate that this pertains to the conditioning and neighboring sites, respectively. In simpler terms, COR indicates the probability an event occurs at both the neighboring and the conditioning sites simultaneously, with "event" defined by the return period thresholds p_i and p_j. Varying these thresholds permits us to examine the similarity between the different H&T models at different degrees of extremity. COR can take any value between 0% and 100%. Figures 4c and 4d show $\mathrm{COR}_{5|5}$ for an example conditioning site and its neighbors for the gauge- and model-driven event sets respectively: In other words, this is the probability a neighbor exceeds the 5-year flow given the conditioning site exceeds the 5-year flow.

The difference between the COR using gauges versus using modeled streamflow at each neighbor can be summarized by the root mean square error at each conditioning site; the co-occurrence rate error (CORE), in units of percentage points (%pts). The CORE is also computed under varying scenarios to better understand model performance:

$$\mathrm{CORE}_{p_j|p_i,\,x} = \sqrt{\frac{\sum_1^N\left[\left(\mathrm{COR}^{obs}_{p_j|p_i} \geq x\right) - \left(\mathrm{COR}^{mod}_{p_j|p_i} \geq x\right)\right]^2}{N}} \tag{3}$$

where N is the number of neighbors to the conditioning site, *obs* and *mod* superscripts indicate whether COR is calculated using the gauge- or model-based events, respectively, and x is a COR threshold above which errors are computed. From Figure 4e, it is evident errors ($\mathrm{COR}^{obs} - \mathrm{COR}^{mod}$) at distal neighbors are low simply because both gauge and model approaches simulate (close to) 0 correlation between such locations and the conditioning site. It may not be particularly discriminatory to reward the model-based method for correctly simulating zero dependence at distal sites, so varying x permits us to examine CORE where conditioning-neighbor correlations are of varying strengths. $\mathrm{CORE}_{5|5,0}$ (5-year event co-occurrence error at all neighbors) for the conditioning site in Figure 4 is 5.1%pts, while $\mathrm{CORE}_{5|5,5}$ (errors at the sites where neighbor co-occurrence >5%) is 9.3%pts.

We further scrutinize model-based H&T error by splitting COREs based on relevant geophysical and climatic characteristics. Each site is grouped within a hydrologic landscape region (based on gauge location) as defined by the USGS (Winter, 2001; Wolock, 2003), which specifies the climatic setting (humid, subhumid, semiarid, and arid), land surface form (plain, plateau, and mountain), and geologic texture (permeable/impermeable soils/bedrock) of river basins across the United States. These hydrologic regions are depicted in Figure S1. Further, the USGS gauge metadata are used to divide sites into three classes of upstream drainage area (small [<1,800 km^2], medium [1,800–7,200 km^2], and large [>7,200 km^2]; each group an equal-population tertile). Since the original COR calculation is pairwise, a given conditioning/neighboring site pair often do not fall within the same category. Hence, COREs are reported as an interaction within or between each grouping for climatology, morphology, geology, and drainage area. For example, the conditioning site in Figure 4 falls in the small drainage area category. $CORE_{5|5,0}$ for this conditioning site and all other small rivers (small-small) is 4.7%pts, for this site and medium rivers (small-medium) is 5.0%pts, and for this site and large rivers (small-large) is 4.8%pts.

To examine the spatial footprint of the synthetic events and how it scales with event magnitude, we perform two comparative tests between the gauge- and model-based event sets. First, where a given conditioning site exceeds a given return period threshold, we compute the median return period across these events experienced at each possible neighbor (based on the spatial limits set out in section 2.2). Assimilating these data for all conditioning sites and plotting median return period against distance from the conditioning site illustrates the distance decay in extremity simulated by the gauge- and model-based approaches. It is here that we expect to see evidence of the aforementioned asymptotic dependence of river flows and the extent to which this differs between data sources. Second, we can examine how the areal extent of flood events change as the conditioning site gets increasingly extreme. Our expectation here is that more extreme events are more widespread in space, while less extreme events are more localized. Again, we are interested in how this relationship differs between H&T model data sources.

For comparisons in a risk-based context, we can plot gauge- and model-based H&T losses on a probability-loss curve. In this instance, probability refers to the empirical event loss quantile; that is, the 99th quantile loss is the loss that has a 1% chance of being exceeded in the event set (the 167,400th most damaging event). Since we have 100 samples of loss for each event—representative of the uncertainty in depth-damage relationships—we can examine how different the probability-loss curves are in the context of this variability.

This suite of analyses, applied to each set of model input data in turn and contrasted with its gauge-based counterpart, can be summarized as follows:

- event co-occurrence rate calculated at each site for different levels of extremity;
- co-occurrence rate isolated for different hydrologic characteristics;
- examination of return period decay with distance when the conditioning site is extreme, and how this scales with increasing extremity;
- examination of event size when the conditioning site is extreme and how this scales with increasing extremity; and
- calculation of flood damages arising from the synthetic event set.

3. Results and Discussion

We can first examine the CORs simulated by the model-based approaches and contrast these with those simulated by the gauge-based H&T model. Figure 5 illustrates $COR_{5|5}$ (an event defined as exceeding the 5-year flow at conditioning and neighboring sites) for all conditioning-neighboring site pair combinations. The total height of the stacked bars (Figures 5a–5c) represent the proportion of model-based CORs that fall within four subjective bins (COR of 0–2%, 2–5%, 5–10%, and 10–100%) relative to the binned distribution of gauge-based CORs. This distribution is selected to ensure large enough sample size across bins which loosely capture a variety of dependence strengths. A bar height of 100% indicates the same number of model and gauge location pairs fall within a particular bin. Most bin heights are close to the 100% optimum, meaning the distribution of site-to-site dependence for all location pairs is broadly similar between model and gauge methods. CaMa-Flood (91%) and PCR-GLOBWB (94%) experience a slight drop in the number of

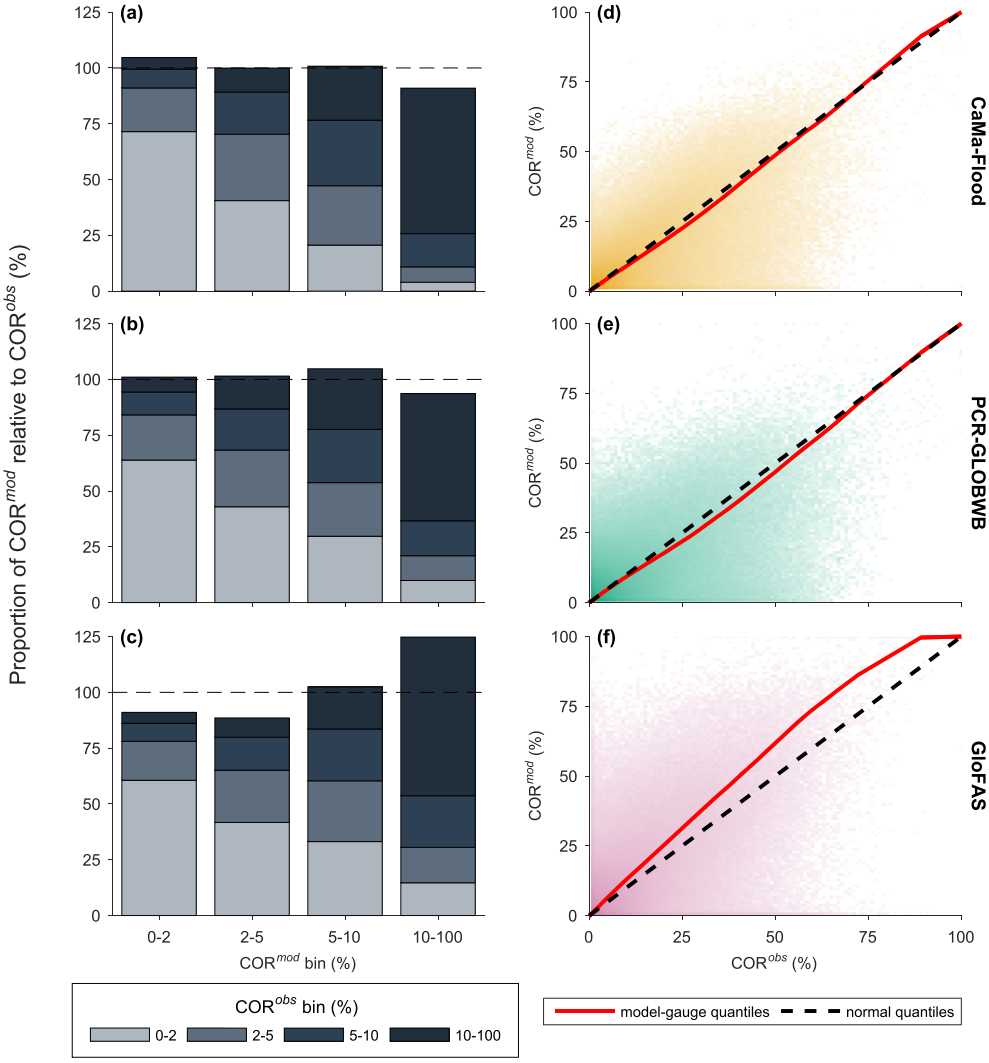

Figure 5. Co-occurrence rate ($COR_{5|5}$) differences between modeled and gauged approaches for all conditioning-neighboring site pairings. Panels (a)–(c) show the proportion of location pairs which have a co-occurrence rate falling within one of four bins according to the model (COR^{mod}) relative to the gauge (COR^{obs}) for CaMa-Flood, PCR-GLOBWB, and GloFAS respectively. The dashed line at $y = 100\%$ indicates an identical number of location pairs within each bin between gauge- and model-based methods. The shaded sections of each bar indicate the proportional constitution of location-level COR^{obs} bins within each COR^{mod} bin. For example, 68% CaMa-Flood 0–2% CORs are also 0–2% gauge-based CORs, while 19%, 8%, and 5% of the CaMaFlood 0–2% bin contains locations pairs with COR^{obs} 2–5%, 5–10%, and 10–100%, respectively. The total number of CaMa-Flood 0–2% CORs is 105% of 0–2% COR^{obs}s. Panels (d)–(f) show COR^{obs} versus COR^{mod} for CaMaFlood, PCR-GLOBWB, and GloFAS, respectively. The colored dots indicate the COR at every location pair, with darker colors signaling a greater density of data points. The red lines are COR^{obs} versus COR^{mod} quantiles, with a perfect match between these illustrated by the dashed line.

high-dependence (COR of 10–100%) locations relative to the gauges, while GloFAS (125%) predicts too many high-dependence sites. Low- to medium-dependence (COR 0–10%) sites are closely replicated by CaMa-Flood and PCR-GLOBWB, but there are too few of them simulated by GloFAS (91% and 89% in 0–2% and 2–5% bins, respectively).

While this represents very high performance versus the gauge-based method *in aggregate*, the colors in each bar stack indicate that, to some degree, nearing the 100% optimum masks some local-level errors. Each discrete color patch corresponds to the COR bin the gauge-based method places a particular location-pair in and thus indicates whether the model approaches have placed individual location-pairs in the same bin as the gauge-based method. The optimum result would be the entirety of the 0–2% COR^{mod} bin bar having

the lightest gray color of the COR^{obs} 0–2% bin (according to the figure legend), the entirety of the 2–5% COR^{mod} bin bar having the slightly darker gray color of the COR^{obs} 2–5% bin, and so on. Most of the model-predicted high-dependence site pairings (COR^{mod} of 10–100%) consist of gauge-predicted high-dependence site pairings (COR^{obs} of 10–100%), yet other COR^{mod} bins also contain 10–100% COR^{obs} bin locations. 72%, 61%, and 57% of the model-predicted high-dependence locations are also gauge-predicted high-dependence locations for CaMa-Flood, PCR-GLOBWB, and GloFAS, respectively. The remaining 10–100% COR^{mod} bin site-pairs are made up of locations where $COR^{obs} \neq$ 10–100%. Model-predicted low-dependence locations (COR^{mod} of 0–2%) are similarly mostly populated by gauge-predicted low-dependence locations (COR^{obs} of 0–2%). 68 percent (CaMa-Flood), 63% (PCR-GLOBWB), and 67% (GloFAS) of model-predicted low-dependence locations are also gauge-predicted low-dependence locations. The central medium-dependence bins (COR^{mod} of 2–10%), however, are clearly populated by a mixture of location-pairs with COR^{obs} values from different bins. Indeed, the 5–10% COR^{mod} bin (for all models) is almost equally constituted of location-pairs with COR^{obs} from all bins: Only 29% (CaMa-Flood), 23% (PCR-GLOBWB), and 23% (GloFAS) of the 5–10% COR^{mod} bins are locations with 5–10% COR^{obs}. Thus, location-pairs which have (subjectively) high or low dependence strengths according to the gauge-based method are generally also classified as such by the models. Location-pairs of (subjectively) medium-dependence strength—which only occasionally experience joint extreme flooding—are misidentified by the models more often.

Figure S2 illustrates that the picture is much the same when event definition is altered to account for return periods of different magnitude at the conditioning site ($COR_{5|20}$, $COR_{5|50}$, and $COR_{5|100}$). CaMa-Flood and PCR-GLOBWB have near-identical binned distributions of pairwise correlation magnitude to the gauge-driven H&T model, with slightly lower proportions of high-dependence locations (COR^{mod} of 10–100% < COR^{obs} of 10–100%). GloFAS slightly underpredicts the number of low-dependence locations (COR^{mod} of 0–10% < COR^{obs} of 0–10%) while overpredicting the number of high-dependence locations (COR^{mod} of 10–100% > COR^{obs} of 10–100%). It appears, then, that this set of results is insensitive to p_i thresholds in the determination of COR. This holds true when each COR^{mod} bar stack is broken down into the COR^{obs} of its component location-pairs also. Low- (0–2%) and high-dependence (10–100%) COR^{obs} locations are generally modeled as such, while medium-dependence COR^{mod} locations (2–10%) are generally a mixture of locations with varying COR^{obs} values.

Overall, at the aggregate nationwide level, overpredictive and underpredictive errors in the characterization of flood dependence seem to cancel out to produce a population of model-based CORs that are broadly similar to gauge-based CORs. This is further evidenced by the plots in Figures 5d–5f, where the bulk of COR^{obs} versus COR^{mod} (darker shading) are clustered around the 1:1 line, albeit with much scatter. The quantiles of the gauge-based CORs versus those of CaMa-Flood and PCR-GLOBWB are a very close fit to quantiles of a normal distribution centered on 0 error. The Q-Q plot of GloFAS illustrates a tendency toward overpredicting location-pair dependence strength, particularly for highly correlated sites.

Figure 6 illustrates the COR error ($CORE_{5|5,0}$) at each conditioning site, where an event is defined as a conditioning-neighboring site pair jointly exceeding their 5-year flow and errors are computed for all locations within the spatial limits. As a whole and for all models, COREs appear relatively low: The majority of errors lie within the 5–10%pts bin (e.g., a location-pair with COR^{obs} of 24% and COR^{mod} of 17% would fall in this group). 21% of CaMa-Flood COREs are <5%pts, 77% are <10%pts. For PCR-GLOBWB, 4% and 70% of COREs are <5%pts and <10%pts, respectively. GloFAS COREs are <5%pts and <10%pts for 4% and 59% of conditioning sites, respectively. Geographic patterns of model skill are similar across all input data sources. Model- and gauge-based CORs seem to converge in the south: around the Arkansas and Ohio Rivers and at sites in Alabama, Mississippi, Tennessee, and Kentucky in southeast United States. COREs increase in the west: most notably across the Rocky Mountains, where COREs exceed 15%pts, and also in New England. Certainly, in the arid Rockies, this is consistent with the quality of the rainfall forcing, where unresolved steep topography can result in an underprediction of extreme precipitation from orographic uplift (Elvidge et al., 2019). Similarly, drylands flooding is mostly due to localized convective storms, meaning hydrological model grid and temporal coarseness may smooth out precipitation extremes. Equally, the proportion of river flow retained in reservoirs or diverted for human use is high in the arid United States, meaning modeled flows may overestimate their true values: Only PCR-GLOBWB attempts to address this where dam operation data are available. Further, by definition, the infrequency of flood events in arid climates

Figure 6. Event co-occurrence rate errors for all sites, where an event is defined as a neighbor and conditioning site concurrently exceeding the 5-year return period. Darker shades of blue indicate a closer match between gauge- and model-generated events. (a) CaMa-Flood, (b) PCR-GLOBWB, and (c) GloFAS.

results in sensitivity of the fitted extreme value distribution tail to relatively few data points. Slight errors in the simulated flow can thus cascade into large deviations in the gradient of the growth curve.

Table S2a isolates these errors ($CORE_{5|5,0}$) for different climatologies. As above, we can see a deterioration in model skill in more arid regions, where hydrological processes are complex, flood response is spatially and temporally heterogeneous, and precipitation data are more uncertain (Beck et al., 2019). Model-driven event simulation involving more humid locations generally replicate the spatial footprint of gauge-driven events more closely. Table S2b illustrates higher model performance on the plains and lower performance in mountainous areas, consistent with results from Figure 6 and known inaccuracies in rainfall products in mountainous areas (Beck et al., 2019). Skill on plateaus lies somewhere in between that of plains and mountains, perhaps due to the diversity of flood drivers in such regions (e.g., snowmelt and ice-jam floods as well as rainfall-driven events; Stein et al., 2019). In Table S2c, model performance appears relatively insensitive to the permeability of the bedrock but improves as soils become more impermeable. This is likely consistent with our lack of knowledge on subsurface processes (Fan, 2019) and the quality of subsurface process

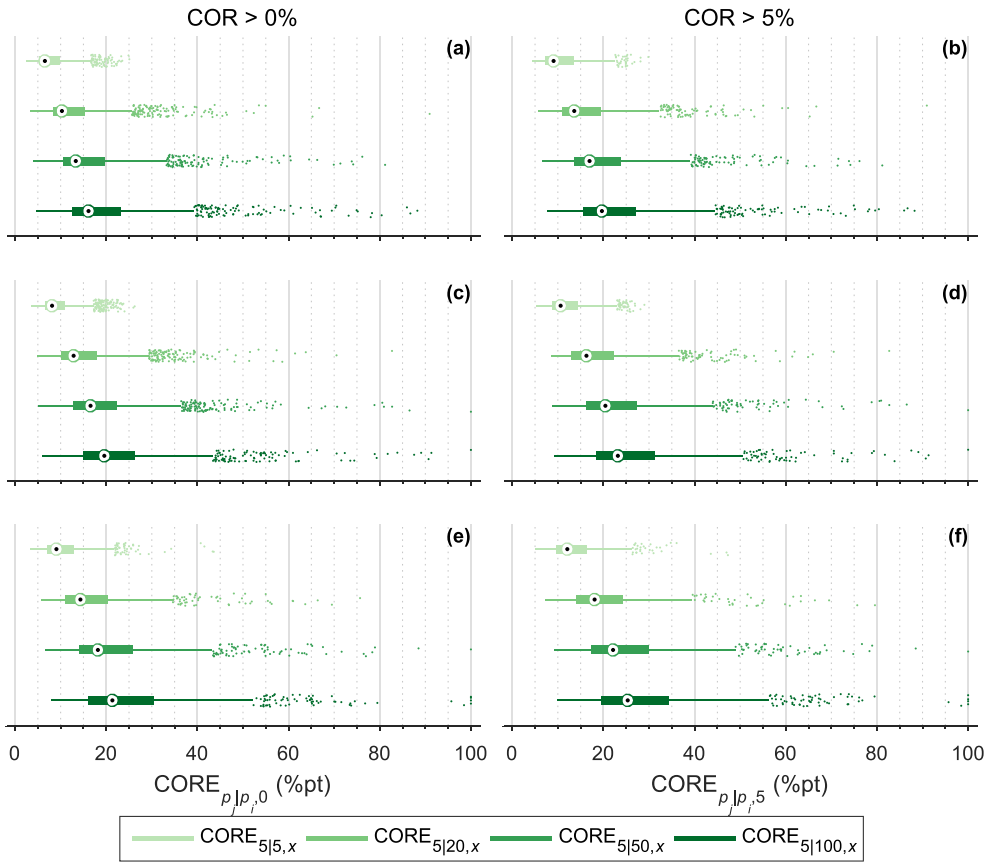

Figure 7. Event co-occurrence rate errors, where events are defined using four different return period thresholds (p_i in Equation 2) and errors are computed at sites with two different strengths of co-occurrence rate (x in Equation 3). Circled dots represent the median CORE, the filled boxes are the interquartile (Q_{25}–Q_{75}) range, and the lines are the whiskers, with outliers represented by dots. Darker shades of green indicate more extreme events at the conditioning site. The left-hand side represent COREs for all sites; the right-hand side represent COREs for sites which are at least 5% correlated. (a, b) CaMa-Flood, (c, d) PCR-GLOBWB, and (e, f) GloFAS.

representation in global hydrological models (Gleeson et al., 2019). Where the translation from rainfall to streamflow is predominantly surface runoff, as in impermeable catchments, modeled event footprints more closely resemble gauge-driven ones. In Table S2d, COREs are stratified by river size (expressed as catchment area). Here, we can see that errors in the characterization of site-to-site correlations consistently decrease as rivers get smaller. This may be due to large river flooding requiring the accurate simulation of the timing of multiple flood waves as they aggregate from tributaries, while smaller river flooding may be mostly controlled by the spatial patterning in the meteorological forcing. Equally, larger rivers are much more likely to be subject to human influence via flow regulation, which are liable to increase COREs where this is unaccounted for.

Overall, this thread of analysis indicates that event co-occurrence errors are sensitive to (in order of decreasing sensitivity) climatology, drainage area, land surface form, soil permeability, and bedrock permeability. More humid regions exhibit higher performance than more arid regions; smaller rivers have higher model performance than larger rivers; performance on flatter terrain is higher than in mountainous areas; events on impermeable soils are better simulated than on permeable soils; and performance is generally insensitive to bedrock permeability. This is consistent across all three global hydrological models studied.

We further explore how model performance changes with different return period and COR thresholds (p_i and x in Equation 3), as $\text{CORE}_{5|5,0}$ may (i) mask errors in the characterization of more extreme events than the 1 in 5 year and (ii) be misrepresented by numerous distal, low-dependence sites which are easier to model as such. Figure 7 displays the same information contained in the histograms of

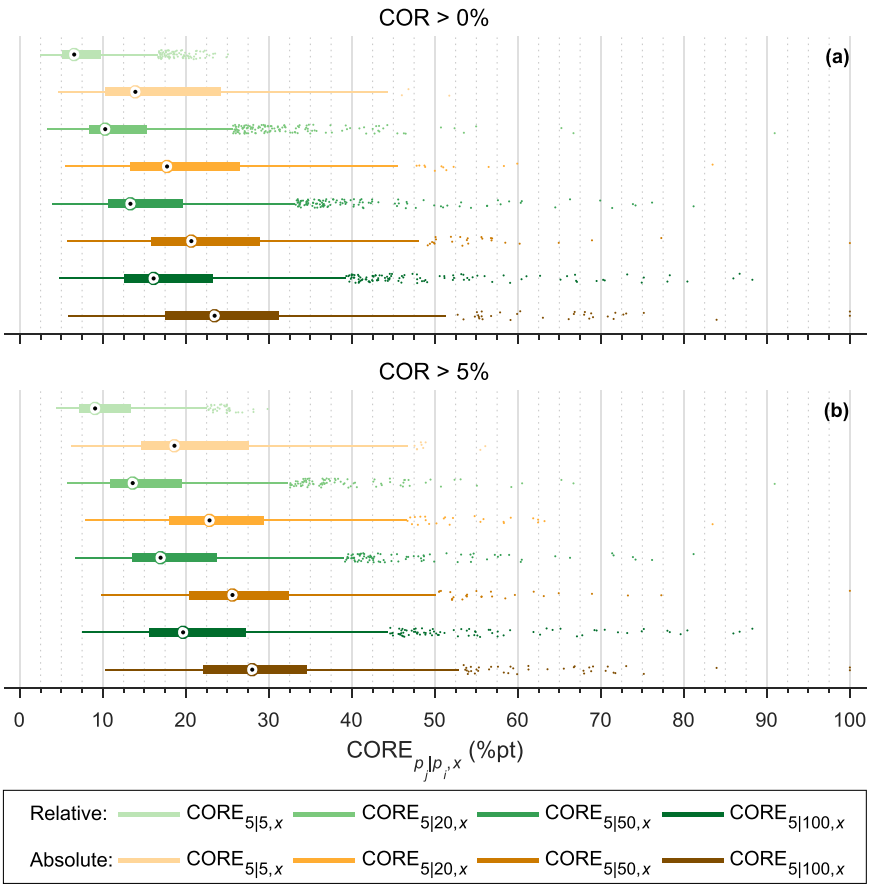

Figure 8. Event co-occurrence rate errors using different thresholds of return period (p_i) and co-occurrence rate (x), computed when COR^{mod} is characterized based on relative flow exceedance (greens) versus when absolute simulated discharges are used instead (oranges). Results are shown for CaMa-Flood: The green plots are identical to those shown in Figures 7a and 7b. Circled dots represent the median CORE, the filled boxes are the interquartile (Q_{25}–Q_{75}) range, the lines are the whiskers, with outliers represented by dots. Darker shading indicates more extreme events at the conditioning site.

Figure 6 but as boxplots with varied p_i and x thresholds. COREs appear to increase as the conditioning site becomes more extreme and as it encompasses the errors of more dependent sites only. CaMa-Flood median $CORE_{5|5,0}$ is 6.5%pts, while $CORE_{5|100,0}$ is 16.1%pts. Removing low-dependence sites from this summary metric, CaMa-Flood median $CORE_{5|5,5}$ and $CORE_{5|100,5}$ are 9.1%pts and 19.7%pts, respectively. This is perhaps as expected, since the spatial averaging effect of coarse meteorological forcing tends to blunt the largest extremes. This is especially true for smaller river basins, whose flood response is governed by smaller, poorly resolved weather events. Even if we are only interested in relative flow exceedances, the effect of this is to flatten the tails of annual flow maxima. Meanwhile, the model-based events appear more skillful when the test involves a more forgiving classification of extreme versus nonextreme flows (where $p_i = 5$). It should also be noted that increasing values of p_i increasingly rely on extrapolations beyond the observation or simulation period of 35 years, meaning the gauge-based H&T benchmark is less reliable in the computation of extremal COREs. The wealth of data simulated by large ensembles of climate models, commonly used in detection and attribution studies, has the potential for addressing the issue of short record length by leveraging many thousands of years of realizations of past weather, rather than relying on a single version of history as we do in this analysis (Mizuta et al., 2017; van der Wiel et al., 2017; van Oldenborgh et al., 2017). Even so, in spite of rising COREs with increasing p_i, characterizing site-to-site correlations within ~20%pts on average even for very rare events may well be acceptable for application in flood risk modeling, where uncertainties related to inundation modeling, exposure geolocation, and flood vulnerability may dominate beyond those involved in generating the event sets.

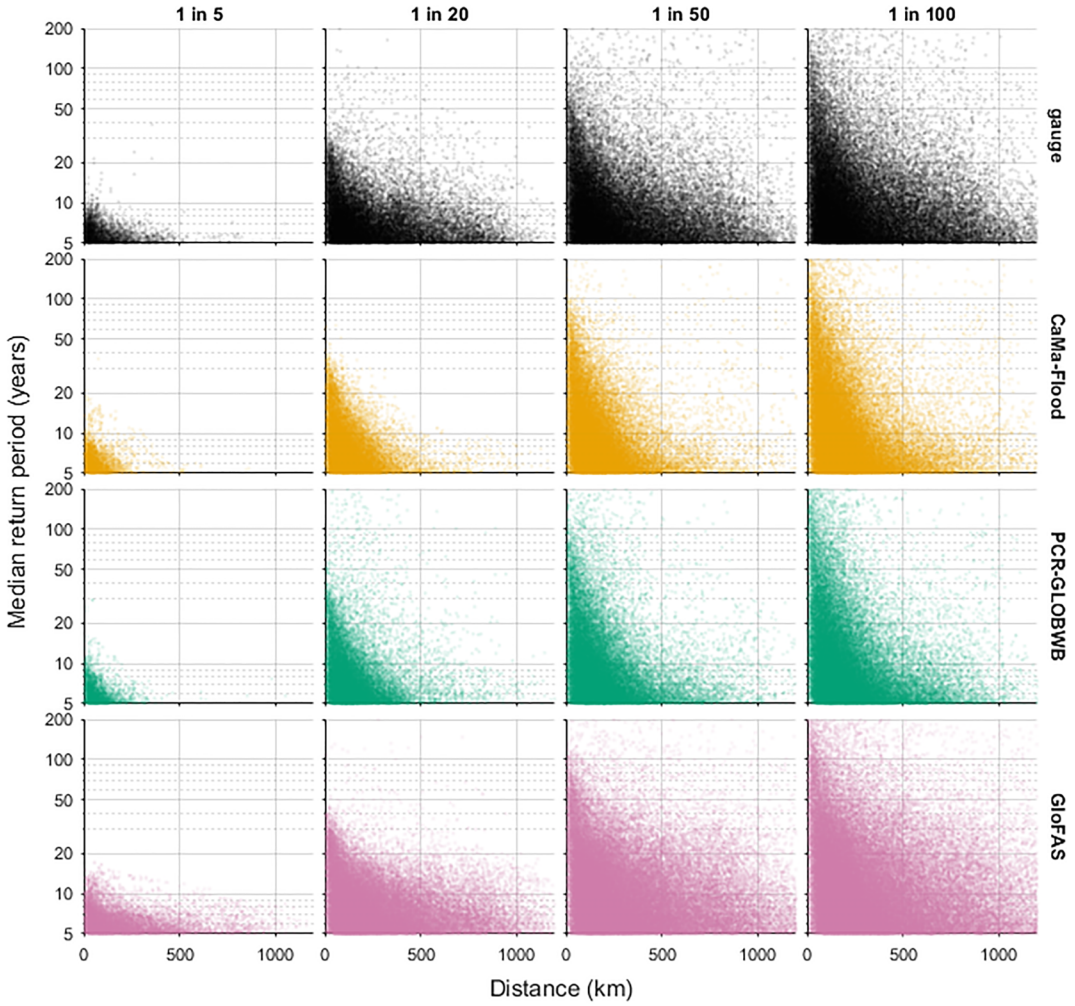

Figure 9. Median return period at a neighboring site versus its distance from the conditioning site, given the conditioning site exceeds the 1 in 5-, 20-, 50-, and 100-year return period (left–right column). Each row corresponds to a separate H&T input data source.

Rather than taking the relative flow exceedances as output, disparities between model- and gauge-based approaches can be considered when the absolute volumetric flows simulated by the hydrological models and synthesized by the H&T model are used instead. The results of this for CaMa-Flood are shown in Figure 8. Return periods for the absolute volumetric discharges simulated by the CaMa-Flood-based events are computed based upon the flood frequency curve of the relevant USGS gauge, while the relative flow exceedance return periods are computed from the flood frequency curve based on the CaMa-Flood simulated discharge time series at the given location. Median absolute co-occurrence rate errors, with events defined fairly loosely ($CORE_{5|5}$), are double relative COREs. $CORE_{5|5,0}$ increases from 6.5 to 13.9%pts (Figure 8a), $CORE_{5|5,5}$ increases from 9.1 to 18.6%pts (Figure 8b). Indeed, across all return period thresholds (p_i) errors increase substantially, to the extent that even the lower quartiles (Q_{25}) of absolute volumetric errors are greater than the median of relative exceedance errors in all cases. The between-site co-occurrence rate during 100-year floods (when COR > 5%) is misestimated by 28.0%pts on average when absolute flows are used, compared to 19.7%pts when relative exceedances are used instead. Although a benchmark for acceptability is difficult to define in terms of CORE, it is clear that the use of absolute volumetric flows simulated by global hydrological models results in a set of synthetic events that are considerably less plausible than if relative flow exceedances are employed.

We can further compare more general characteristics of the model- and gauge-based event sets through examinations of the size of simulated events and their extremity. In Figure 9, gauge-based H&T events

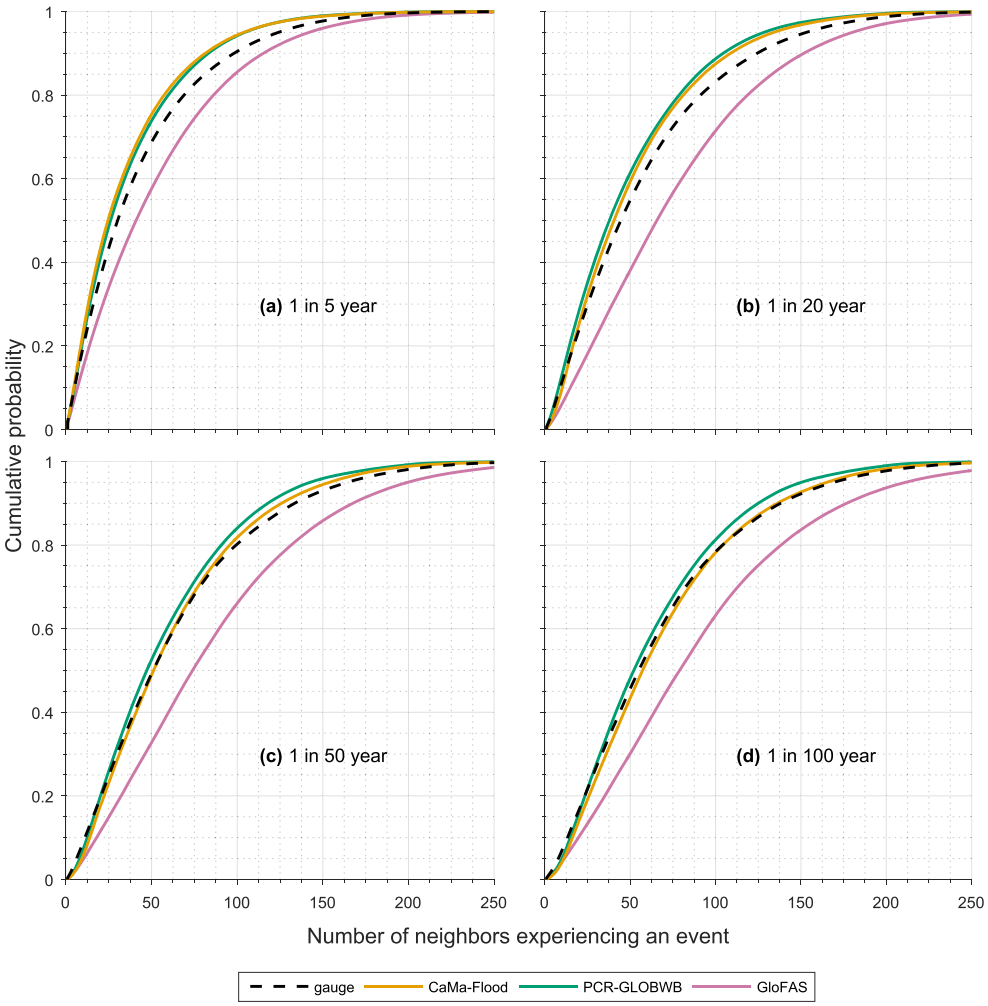

Figure 10. Distribution of synthetic event sizes when the conditioning site exceeds the (a) 1 in 5-year, (b) 1 in 20-year, (c) 1 in 50-year, and (d) 1 in 100-year return period for all H&T input data. An event at a neighboring site is defined as where it exceeds the 1 in 5-year return period. Table S3 summarizes the data in this graphic.

show a decay of return period with distance from the conditioning site. As the conditioning site becomes more extreme, extreme events generally occur at greater distance from the conditioning site, but the most extreme return periods during such events remain more localized (illustrated by the sharper decline in return period with distance for more extreme conditioning site events). These phenomena are generally well replicated by the hydrological model-based H&T events. CaMa-Flood and PCR-GLOBWB generally show sharper declines in return period with distance than the gauge-based approach, meaning distal sites become less extreme too quickly or events are generally too localized. Conversely, the distance decay of return period illustrated by GloFAS is gentler than that of the USGS gauges. GloFAS events are generally too widespread, where extreme return periods are experienced at greater distances from conditioning sites than for events simulated used river gauge data. One possible cause of this may be the lack of a floodplain inundation component in GloFAS, which would serve to dampen the flood wave peak. CaMa-Flood and PCR-GLOBWB, where synthetic event footprints are not as widespread, have some representation of inundation in their model structures. Equally, the effect of GloFAS calibration (CaMa-Flood and PCR-GLOBWB are uncalibrated) may have unintentionally altered the correlation of relative flow exceedances, in spite of improving the correspondence between observed and modeled absolute discharges.

The variation in the number of sites experiencing an event when the conditioning site exceeds certain return periods is illustrated for the different input data sources in Figure 10 and Table S3, where observations

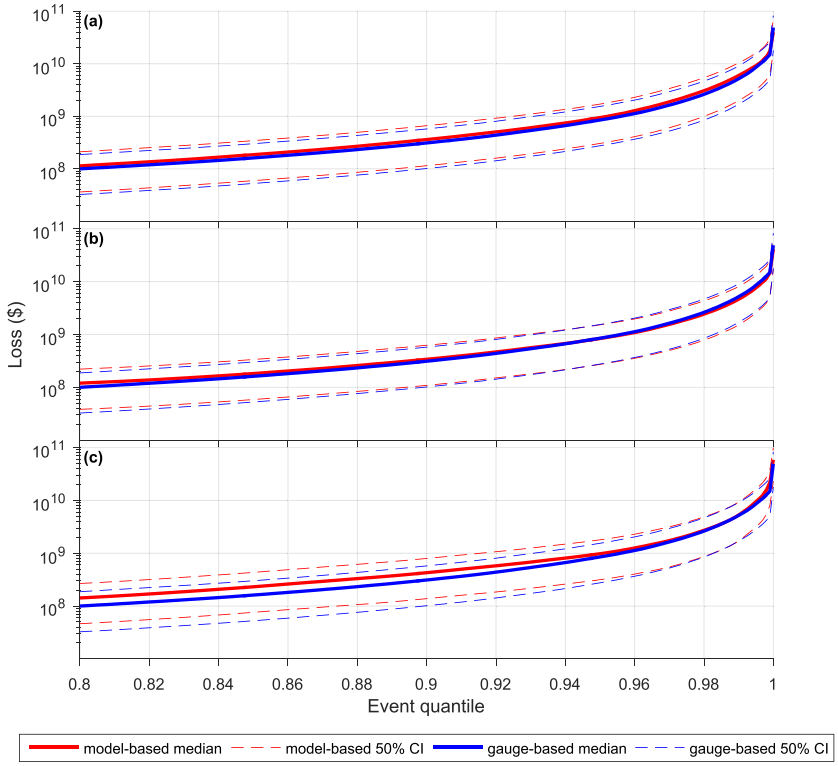

Figure 11. Distribution of losses arising from the most damaging 20% of flood events when calculated using the gauge-based (blue) versus model-based (red) synthetic event set. Dashed lines represent the 50% (Q_{25}–Q_{75}) confidence interval (CI) given vulnerability uncertainty. (a) CaMa-Flood, (b) PCR-GLOBWB, and (c) GloFAS.

made in light of Figure 9 are generally reaffirmed. Gauge-based events are generally more widespread as their return period increases. GloFAS-based events impact too many sites across all return periods, while CaMa-Flood and PCR-GLOBWB are a closer match to the gauge-based events (yet impact too few sites). For instance, from Table S3, 22% of gauge-based 100-year events impact at least 100 sites, while for CaMa-Flood, PCR-GLOBWB, and GloFAS, the event proportions are 22%, 19%, and 37%, respectively. For 5-year events, the proportions of events which impact at least 100 sites are 10%, 6%, 6%, and 15% for gauge, CaMa-Flood, PCR-GLOBWB, and GloFAS, respectively.

On the whole, the relationships between flood event probability and spatial extent exhibited by the gauge-based approach are well replicated by the model-based methods. Return period decay with distance —and the scaling of this decay with increasing extremity—are similar across input data sources. Equally, the extent of flood events given their return period—and how the extent increases with increasing extremity—appear consistent too.

The final step in the comparison of synthetic events generated by different input flow data involves a comparison of losses when these events are cascaded through a hydraulic model to produce inundation depths arising from each event, which, in turn, are intersected with building and vulnerability information to generate losses. The vulnerability functions are probabilistic, meaning they relate a given inundation depth to an array of possible flood damages (following Wing et al., 2020). As a result, each synthetic event has a probability distribution of possible damages: The upper and lower quartiles of this are plotted in Figure 11 (the most extreme 20% of events are shown). Comparing model- and gauge-based approaches, losses are virtually indistinguishable for CaMa-Flood and PCR-GLOBWB. Even for GloFAS, the median event loss never falls outside of the 50% confidence interval, though the assertion of events being too large is evident in the event-loss curve. For the most damaging events, losses appear to converge. This perhaps indicates that the largest possible flood events are captured by the behavior inherent to simulated flow time series or that there is some constraint in the physical reality of a synthetic event imparted by the hydraulics.

From Table S4, the 99th quantile event—one of the most extreme simulated—generates an average of $5.4 billion of damage according to the gauge-based approach, but the central 50% of losses given vulnerability uncertainty ranges between $1.8 and $9.5 billion. The median 99th quantile loss from model-based approaches are $6.2, $4.8, and $5.6 billion, while the central 50% of losses are $2.1–10.8, $1.6–8.5, and $1.8–9.8 billion, for CaMa-Flood, PCR-GLOBWB, and GloFAS, respectively. It is evident, then, that all flow data sources when used as input to a stochastic flood risk model produce losses to a nationwide inventory of U.S. buildings that are within the uncertainty inherent in the characterization of flood vulnerability. While this uncertainty is very large, preceding tests incorporating only the physical modeling further evidence the correspondence between model- and gauge-driven approaches.

4. Conclusions

The proliferation of global hydrological models in recent years has enabled a step change in our understanding of the response of the terrestrial water cycle to meteorology and climate, though their defensible application in high-resolution flood risk frameworks which capture the spatial structure of real flood events has yet to be evidenced. In this study, we use simulated and observed discharge time series as input to a U.S.-wide conditional exceedance model based on Heffernan and Tawn (2004) and Quinn et al. (2019) to generate tens of millions of plausible extreme flood events. We find that, while the absolute simulated extreme flows may be unreliable for global hydrological models, the relative simulated flow exceedances largely resemble flood events based on observations from gauge data. Using a relatively loose definition of "an event footprint" (sites experiencing >5-year flow simultaneously), errors in the quantification of site-to-site correlations are generally less than 10%pts. Where events are defined more specifically (e.g., sites experiencing >100-year flow), errors are generally below 20%pts for all return periods. Co-occurrence quantification errors are generally higher for larger rivers, in mountainous areas, and in arid climates, while errors are lower for small streams, on plains, and in humid climates. These location-level errors are mostly mitigated (positive and negative biases cancel) in aggregate, as evidenced by the similar number of location-pairs which fall within subjective bins describing the magnitude of their correlation across all input flow data sources. The relationships between extremity and event size exhibited by the gauge-generated flood events are also broadly replicated by the model-based approaches. In their application within a catastrophe modeling framework to a portfolio of >100M buildings, risk quantifications are largely indistinguishable regardless of the input flow data source. To reiterate, any of these global hydrological models appear fit-for-purpose in this context.

It is important to note that conclusions drawn over the United States may not be directly transferrable globally. The climate reanalysis forcing these hydrological models is likely of higher quality over the United States than globally, since the richness of meteorological observations here will have been assimilated into the global product. Further, the effect of model calibration to streamflow observations may also limit the applicability of these results globally, since the United States has a dense network of river gauges with which to tune model parameters. That said, both CaMa-Flood and PCR-GLOBWB are uncalibrated and generally outperformed the calibrated GloFAS in this analysis of spatial dependence characterization. This is, perhaps, a promising conclusion for the prospect of obtaining comparable skill in the majority of world regions where rigorous calibration is not possible: Event footprints may be largely determined by the climate and topology of the river network rather than model parameterization. It is also worth noting that these U.S.-only results may, in some cases, be an understatement of global hydrological model skill in the context of spatial correlation. U.S. rivers are among the most intensively engineered and managed in the world. The presence and operation of dams and other flow control structures presents a formidable modeling challenge, especially for a global hydrological model. Rivers in other parts of the globe (certainly in the developing world), in general, have flow regimes much more akin to their natural behavior and thus may be simulated more accurately by global models that lack these local anthropogenic details. Further, the United States consists of diverse climatological regions with complex hydrological responses to meteorological events, including the arid and semiarid western United States; topographically complex Rocky, Appalachian, Sierra Nevada, and Cascade mountain ranges; snow-dominated regions in the Rockies and the northern United States; atmospheric river-induced rainfall in the west; and hurricane-induced rainfall in the southeast and east. These characteristics mean many parts of the United States are highlighted as areas of significant bias in evaluations of global climate forcing data (Beck et al., 2017), perhaps indicating some other regions

of the globe may be modeled with greater fidelity. In summary, it is likely that different results may be obtained in different regions of the world, and we cannot yet quantify how representative these U.S. results may be of global performance.

To this end, a replication of this analysis in similarly well-instrumented Europe will be a crucial addition to the results presented here. Equally, examinations outside of Europe and North America where suitable river gauge data are available will constrain the accuracy of such an approach amidst true discharge data paucity. Further work must also examine a computationally tractable discretization of sampling points from hydrologically modeled flow time series. In this analysis—for the purposes of intercomparison—time series were sampled from USGS gauge locations. For operational use in a global catastrophe model, sampling for integration into a H&T-like model will need to be at locations which minimize the heterogeneity of return period upstream while maintaining computational feasibility for the calculation of pairwise dependence. The results from this analysis indicate, for the first time, the suitability of global hydrological models in the characterization of extreme flow dependence. This paves the way for the development of global stochastic flood risk models, which properly simulate and capture the physical structure of real events, to more accurately constrain the risk of global populations and assets to extreme flooding.

Data Availability Statement

USGS river gauge data are available from https://waterdata.usgs.gov/nwis/. CaMa-Flood model output can be accessed via http://hydro.iis.u-tokyo.ac.jp/~yamadai/cama-flood/. PCR-GLOBWB model scripts are available from https://github.com/UU-Hydro/PCR-GLOBWB_model, with standard parameterization and meteorological forcing available from https://opendap.4tu.nl/thredds/catalog/data2/pcrglobwb/version_2019_11_beta/pcrglobwb2_input/catalog.html, while the model output is available from Sutanudjaja et al. (2018). GloFAS model output can be accessed from the JRC Data Catalog (https://data.jrc.ec.europa.eu/collection/id-00288). The flood inundation maps and stochastic modeling framework are available from Wing et al. (2017) and Quinn et al. (2019). Vulnerability functions are available from Wing et al. (2020).

Acknowledgments
Oliver Wing and Paul Bates were supported by EPSRC Grant EP/R511663/1. Paul Bates was also supported by a Leverhulme Research Fellowship and a Royal Society Wolfson Research Merit award. Jeff Neal was supported by NERC Grant NE/S003061/1. Dai Yamazaki was supported by MEXT Tougou Project JPMXD0717935457. The PCR-GLOBWB simulation was carried out on the Dutch National e-Infrastructure with the support of SURF Cooperative.

References

Alfieri, L., Feyen, L., Salamon, P., Thielen, J., Bianchi, A., Dottori, F., & Burek, P. (2016). Modelling the socio-economic impact of river floods in Europe. *Natural Hazards and Earth System Sciences*, *16*(6), 1401–1411. https://doi.org/10.5194/nhess-16-1401-2016

Alfieri, L., Lorini, V., Hirpa, F. A., Harrigan, S., Zsoter, E., Prudhomme, C., & Salamon, P. (2020). A global streamflow reanalysis for 1980–2018. *Journal of Hydrology X*, *6*, 100049. https://doi.org/10.1016/j.hydroa.2019.100049

Alfieri, L., Salamon, P., Bianchi, A., Neal, J., Bates, P., & Feyen, L. (2014). Advances in pan-European flood hazard mapping. *Hydrological Processes*, *28*(13), 4067–4077. https://doi.org/10.1002/hyp.9947

Allen, G. H., David, C. H., Andreadis, K. M., Hossain, F., & Famiglietti, J. S. (2018). Global estimates of river flow wave travel times and implications for low-latency satellite data. *Geophysical Research Letters*, *45*, 7551–7560. https://doi.org/10.1029/2018GL077914

Bates, P. D., Horritt, M. S., & Fewtrell, T. J. (2010). A simple inertial formulation of the shallow water equations for efficient two-dimensional flood inundation modelling. *Journal of Hydrology*, *387*(1–2), 33–45. https://doi.org/10.1016/j.jhydrol.2010.03.027

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Dutra, E., Fink, G., Orth, R., & Schellekens, J. (2017). Global evaluation of runoff from 10 state-of-the-art hydrological models. *Hydrology and Earth System Sciences*, *21*(6), 2881–2903. https://doi.org/10.5194/hess-21-2881-2017

Beck, H. E., Wood, E. F., Pan, M., & Fisher, C. K. (2019). MSWEP v2 global 3-hourly 0.1° precipitation: Methodology and quantitative assessment. *Bulletin of the American Meteorological Society*, *100*(3), 473–500. https://doi.org/10.1175/BAMS-D17-0138.1

Beven, K. J., & Cloke, H. L. (2012). Comment on "Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water" by Eric F Wood et al. *Water Resources Research*, *48*, W01801. https://doi.org/10.1029/2011WR010982

Blöschl, G., & Sivapalan, M. (1997). Process controls on regional flood frequency: Coefficient of variation and basin scale. *Water Resources Research*, *33*(12), 2967–2980. https://doi.org/10.1029/97WR00568

Bosmans, J. H. C., van Beek, L. P. H., Sutanudjaja, E. H., & Bierkens, M. F. P. (2017). Hydrological impacts of global land cover change and human water use. *Hydrology and Earth System Sciences*, *21*(11), 5603–5626. https://doi.org/10.5194/hess-21-5603-2017

Candogan Yossef, N., Winsemius, H., Weerts, A., van Beek, R., & Bierkens, M. F. P. (2013). Skill of a global seasonal streamflow forecasting system, relative roles of initial conditions and meteorological forcing. *Water Resources Research*, *49*, 4687–4699. https://doi.org/10.1002/wrcr.20350

Coxon, G., Freer, J., Westerberg, I. K., Wagener, T., Woods, R., & Smith, P. J. (2015). A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations. *Water Resources Research*, *51*, 5531–5546. https://doi.org/10.1002/2014WR016532

de Almeida, G. A. M., Bates, P., Freer, J. E., & Souvignet, M. (2012). Improving the stability of a simple formulation of the shallow water equations for 2-D flood modeling. *Water Resources Research*, *48*, W05528. https://doi.org/10.1029/2011WR011570

de Graff, I. E. M., Gleeson, T., van Beek, L. P. H., Sutanudjaja, E. H., & Bierkens, M. F. P. (2019). Environmental flow limits to global groundwater pumping. *Nature*, *574*(7776), 90–94. https://doi.org/10.1038/s41586-019-1594-4

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

Di Baldassarre, G., & Montanari, A. (2009). Uncertainty in river discharge observations: A quantitative analysis. *Hydrology and Earth System Sciences*, *13*(6), 913–921. https://doi.org/10.5194/hess-13-913-2009

Diederen, D., Liu, Y., Gouldby, B., Diermanse, F., & Vorogushyn, S. (2019). Stochastic generation of spatially coherent river discharge peaks for continental event-based flood risk assessment. *Natural Hazards and Earth System Sciences*, *19*(5), 1041–1053. https://doi.org/10.5194/nhess-19-1041-2019

Dottori, F., Salamon, P., Bianchi, A., Alfieri, L., Hirpa, F. A., & Feyen, L. (2016). Development and evaluation of a framework for global flood hazard mapping. *Advances in Water Resources*, *94*, 87–102. https://doi.org/10.1016/j.advwatres.2016.05.002

Elvidge, A. D., Sandu, I., Wedi, N., Vosper, S. B., Zadra, A., Boussetta, S., et al. (2019). Uncertainty in the representation of orography in weather and climate models and implications for parameterized drag. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2567–2585. https://doi.org/10.1029/2019MS001661

Falter, D., Dung, N. V., Vorogushyn, S., Schröter, K., Hundecha, Y., Kreibich, H., et al. (2016). Continuous, large-scale simulation model for flood risk assessments: Proof-of-concept. *Journal of Flood Risk Management*, *9*(1), 3–21. https://doi.org/10.1111/jfr3.12105

Falter, D., Schröter, K., Dung, N. V., Vorogushyn, S., Kreibich, H., Hundecha, Y., et al. (2015). Spatially coherent flood risk assessment based on long-term continuous simulation with a coupled model chain. *Journal of Hydrology*, *524*, 182–193. https://doi.org/10.1016/j.jhydrol.2015.02.021

Fan, Y. (2019). Are catchments leaky? *WIREs Water*, *6*(6), e1386. https://doi.org/10.1002/wat2.1386

Ghizzoni, T., Roth, G., & Rudari, R. (2012). Multisite flooding hazard assessment in the Upper Mississippi River. *Journal of Hydrology*, *412-413*, 101–113. https://doi.org/10.1016/j.jhydrol.2011.06.004

Gilles, D., Young, N., Schroeder, H., Piotrowski, J., & Chang, Y.-J. (2012). Inundation mapping initiatives of the Iowa Flood Center: Statewide coverage and detailed urban flooding analysis. *Water*, *4*(1), 85–106. https://doi.org/10.3390/w4010085

Gleeson, T., Wagener, T., Cuthbert, M., Rahman, S., Bierkens, M. F. P., Döll, P., et al. (2019) Groundwater representation in continental to global hydrologic models: A call for open and holistic evaluation, conceptualization and classification. EarthArXiv preprints. https://doi.org/10.31226/osf.io/zxyku

Grimaldi, S., Petroselli, A., Arcangeletti, E., & Nardi, F. (2013). Flood mapping in ungauged basins using fully continuous hydrologic-hydraulic modelling. *Journal of Hydrology*, *487*, 39–47. https://doi.org/10.1016/j.jhydrol.2013.02.023

Grimaldi, S., Schumann, G. J.-P., Shokri, A., Walker, J. P., & Pauwels, V. R. N. (2019). Challenges, opportunities, and pitfalls for global coupled hydrologic-hydraulic modeling of floods. *Water Resources Research*, *55*, 5277–5300. https://doi.org/10.1029/2018WR024289

Haberlandt, U., & Radtke, I. (2014). Hydrological model calibration for derived flood frequency analysis using stochastic rainfall and probability distributions of peak flows. *Hydrology and Earth System Sciences*, *18*(1), 353–365. https://doi.org/10.5194/hess-18-353-2014

Hannah, D. M., Demuth, S., van Lanen, H. A. J., Looser, U., Prudhomme, C., Rees, G., et al. (2011). Large-scale river flow archives: Importance, current status and future needs. *Hydrological Processes*, *25*(7), 1191–1200. https://doi.org/10.1002/hyp.7794

Harris, I., Jones, P. D., Osborn, T. J., & Lister, D. H. (2014). Updated high-resolution grids of monthly climatic observations—The CRU TS3.10 dataset. *International Journal of Climatology*, *34*(3), 623–642. https://doi.org/10.1002/joc.3711

Heffernan, J. E., & Tawn, J. A. (2004). A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *66*(3), 497–546. https://doi.org/10.1111/j.1467-9868.2004.02050.x

Herzbach, H., de Rosnay, P., Bell, B., Schepers, D., Simmons, A., Soci, C., et al. (2018). Operational global reanalysis: Progress, future directions and synergies with NWP, *ERA Report Series 27*, Reading, UK: European Centre for Medium Range Weather Forecasts. https://doi.org/10.21957/tkic6g3wm

Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., et al. (2013). Global flood risk under climate change. *Nature Climate Change*, *3*(9), 816–821. https://doi.org/10.1038/nclimate1911

Keef, C., Svensson, C., & Tawn, J. A. (2009). Spatial dependence in extreme rivers flows and precipitation for Great Britain. *Journal of Hydrology*, *378*(3–4), 240–252. https://doi.org/10.1016/j.jhydrol.2009.09.026

Keef, C., Tawn, J. A., & Lamb, R. (2013). Estimating the probability of widespread flood events. *Environmetrics*, *24*(1), 13–21. https://doi.org/10.1002/env.2190

Kendon, E. J., Ban, N., Roberts, N. M., Fowler, H. J., Roberts, M. J., Chan, S. C., et al. (2017). Do convection-permitting regional climate models improve projections of future precipitation change? *Bulletin of the American Meteorological Society*, *98*(1), 79–93. https://doi.org/10.1175/BAMS-D-15-0004.1

Kidd, C., Bauer, P., Turk, J., Huffman, G. J., Joyce, R., Hsu, K.-L., & Braithwaite, D. (2012). Intercomparison of high-resolution precipitation products over northwest Europe. *Journal of Hydrometeorology*, *13*(1), 67–83. https://doi.org/10.1175/JHM-D-11-042.1

Lamb, R., Keef, C., Tawn, J., Laeger, S., Meadowcroft, I., Surendran, S., et al. (2010). A new method to assess the risk of local and widespread flooding on rivers and coasts. *Journal of Flood Risk Management*, *3*(4), 323–336. https://doi.org/10.1111/j.1753-318X.2010.01081.x

Lehner, B., & Grill, G. (2013). Global river hydrography and network routing: Baseline data and new approaches to study the world's large river systems. *Hydrological Processes*, *27*(15), 2171–2186. https://doi.org/10.1002/hyp.9740

McMillan, H., Krueger, T., & Freer, J. (2012). Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality. *Hydrological Processes*, *26*(26), 4078–4111. https://doi.org/10.1002/hyp.9384

Meigh, J. R., Farquharson, F. A. K., & Sutcliffe, J. V. (1997). A worldwide comparison of regional flood estimation methods and climate. *Hydrological Sciences Journal*, *42*(1), 1–14. https://doi.org/10.1080/02626669709492002

Merz, B., & Blöschl, G. (2009). A regional analysis of event runoff coefficients with respect to climate and catchment characteristics in Austria. *Water Resources Research*, *45*, W01405. https://doi.org/10.1029/2008WR007163

Metin, A. D., Dung, N. V., Schröter, K., Vorogushyn, S., Guse, B., Kreibich, H., & Merz, B. (2020). The role of spatial dependence for large-scale flood risk estimation. *Natural Hazards and Earth System Sciences*, *20*(4), 967–979. https://doi.org/10.5194/nhess-20-967-2020

Mizuta, R., Murata, A., Ishii, M., Shiogama, H., Hibino, K., Mori, N., et al. (2017). Over 5000 years of ensemble future climate simulations by 60-km global and 20-km regional atmospheric models. *Bulletin of the American Meteorological Society*, *98*(7), 1383–1398. https://doi.org/10.1175/BAMS-D-16-0099.1

Neal, J., Keef, C., Bates, P., Beven, K., & Leedal, D. (2013). Probabilistic flood risk mapping including spatial dependence. *Hydrological Processes*, *27*(9), 1349–1363. https://doi.org/10.1002/hyp.9572

Northrop, P. J., Attalides, N., & Jonathan, P. (2017). Cross-validatory extreme value threshold selection and uncertainty with application to ocean storm severity. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, *66*(1), 93–120. https://doi.org/10.1111/rssc.12159

Pappenberger, F., Dutra, E., Wetterhall, F., & Cloke, H. L. (2012). Deriving global flood hazard maps of fluvial floods through a physical model cascade. *Hydrology and Earth System Sciences*, *16*(11), 4143–4156. https://doi.org/10.5194/hess-16-4143-2012

Prudhomme, C., Parry, S., Hannaford, J., & Clark, D. B. (2011). How well do large-scale models reproduce regional hydrological extremes in Europe? *Journal of Hydrometeorology*, *12*(6), 1181–1204. https://doi.org/10.1175/2011JHM1387.1

Quinn, N., Bates, P. D., Neal, J., Smith, A., Wing, O., Sampson, C., et al. (2019). The spatial dependence of flood hazard and risk in the United States. *Water Resources Research*, *55*, 1890–1911. https://doi.org/10.1029/2018WR024205

Salinas, J. L., Laaha, G., Rogger, M., Parajka, J., Viglione, A., Sivalapan, M., & Blöschl, G. (2013). Comparative assessment of predictions in ungauged basins—Part 2: Flood and low flow studies. *Hydrology and Earth System Sciences*, *17*(7), 2637–2652. https://doi.org/10.5194/hess-17-2637-2013

Sampson, C. C., Fewtrell, T. J., O'Loughlin, F., Pappenberger, F., Bates, P. D., Freer, J. E., & Cloke, H. L. (2014). The impact of uncertain precipitation data on insurance loss estimates using a flood catastrophe model. *Hydrology and Earth System Sciences*, *18*(6), 2305–2324. https://doi.org/10.5194/hess-18-2305-2014

Sampson, C. C., Smith, A. M., Bates, P. D., Neal, J. C., Alfieri, L., & Freer, J. E. (2015). A high-resolution global flood hazard model. *Water Resources Research*, *51*, 7358–7381. https://doi.org/10.1002/2015WR016954

Schellekens, J., Dutra, E., Martínez-de la Torre, A., Balsamo, G., Van Dijk, A., Sperna Weiland, F., et al. (2017). A global water resources ensemble of hydrological models: The eartH2Observe Tier-1 dataset. *Earth System Science Data*, *9*(2), 389–413. https://doi.org/10.5194/essd-9-389-2017

Smith, A., Sampson, C., & Bates, P. (2015). Regional flood frequency analysis at the global scale. *Water Resources Research*, *51*, 539–553. https://doi.org/10.1002/2014WR015814

Smith, J. A., Bradley, A. A., & Baeck, M. L. (1994). The space-time structure of extreme storm rainfall in the southern plains. *Journal of Applied Meteorology and Climatology*, *33*(12), 1402–1417. https://doi.org/10.1175/1520-0450(1994)033<1402:TSSOES>2.0.CO;2

Smith, J. A., Villarini, G., & Baeck, M. L. (2011). Mixture distributions and the hydroclimatology of extreme rainfall and flooding in the eastern United States. *Journal of Hydrometeorology*, *12*(2), 294–309. https://doi.org/10.1175/2010JHM1242.1

Sperna Weiland, F. C., van Beek, L. P. H., Kwadijk, J. C. J., & Bierkens, M. F. P. (2010). The ability of a GCM-forced hydrological model to reproduce global discharge variability. *Hydrology and Earth System Sciences*, *14*(8), 1595–1621. https://doi.org/10.5194/hess-14-1595-2010

Sperna Weiland, F. C., van Beek, L. P. H., Kwadijk, J. C. J., & Bierkens, M. F. P. (2012). Global patterns of change in discharge regimes for 2100. *Hydrology and Earth System Sciences*, *16*(4), 1047–1062. https://doi.org/10.5194/hess-16-1047-2012

Stein, L., Pianosi, F., & Woods, R. (2019). Event-based classification for global study of river flood generating processes. *Hydrological Proccesses*, *34*(7), 1514–1529. https://doi.org/10.1002/hyp.13678

Sutanudjaja, E. H., van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H. C., Drost, N., et al. (2018). PCR-GLOBWB 2: A 5 arcmin global hydrological and water resources model. *Geoscientific Model Development*, *11*(6), 2429–2453. https://doi.org/10.5194/gmd-11-2429-2018

Thieken, A. H., Apel, H., & Merz, B. (2015). Assessing the probability of large-scale flood loss events: A case study for the river Rhine, Germany. *Journal of Flood Risk Management*, *8*(3), 247–262. https://doi.org/10.1111/jfr3.12091

Uppala, S. M., Kållberg, P. W., Simmons, A. J., Andrae, U., Da Costa Bechtold, V., Fiorino, M., et al. (2005). The ERA-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society*, *131*(612), 2961–3012. https://doi.org/10.1256/qj.04.176

van Beek, L. P. H., Wada, Y., & Bierkens, M. F. P. (2011). Global monthly water stress: 1. Water balances and water availability. *Water Resources Research*, *47*, W07517. https://doi.org/10.1029/2010WR009791

van der Knijff, J. M., Younis, J., & de Roo, A. P. J. (2010). LISFLOOD: A GIS-based distributed model for river basin scale water balance and flood simulation. *International Journal of Geographical Information Science*, *24*(2), 189–212. https://doi.org/10.1080/13658810802549154

van der Wiel, K., Kapnick, S. B., van Oldenborgh, G. J., Whan, K., Philip, S., Vecchi, G. A., et al. (2017). Rapid attribution of the August 2016 flood-inducing extreme precipitation in south Louisiana to climate change. *Hydrology and Earth System Sciences*, *21*(2), 897–921. https://doi.org/10.5194/hess-21-897-2017

van Oldenborgh, G. J., van der Wiel, K., Sebastian, A., Singh, R., Arrighi, J., Otto, F., et al. (2017). Attribution of extreme rainfall from Hurricane Harvey, August 2017. *Environmental Research Letters*, *12*(12), 124,009. https://doi.org/10.1088/1748-9326/aa9ef2

Vorogushyn, S., Bates, P. D., de Bruijn, K., Castellarin, A., Kreibich, H., Priest, S., et al. (2018). Evolutionary leap in large-scale flood risk assessment needed. *WIREs Water*, *5*(2), e1266. https://doi.org/10.1002/wat2.1266

Wanders, N., & Wada, Y. (2015). Human and climate impacts on the 21st century hydrological drought. *Journal of Hydrology*, *526*, 208–220. https://doi.org/10.1016/j.jhydrol.2014.10.047

Ward, P. J., Jongman, B., Sperna Weiland, F., Bouwman, A., van Beek, R., Bierkens, M. F. P., et al. (2013). Assessing flood risk and the global scale: Model setup, results, and sensitivity. *Environmental Research Letters*, *8*(4), 044019. https://doi.org/10.1088/1748-9326/8/4/044019

Wing, O. E. J., Bates, P. D., Neal, J. C., Sampson, C. C., Smith, A. M., Quinn, N., et al. (2019). A new automated method for improved flood defense representation in large-scale hydraulic models. *Water Resources Research*, *55*, 11,007–11,034. https://doi.org/10.1029/2019WR025957

Wing, O. E. J., Bates, P. D., Sampson, C. C., Smith, A. M., Johnson, K. A., & Erickson, T. A. (2017). Validation of a 30 m resolution flood hazard model of the conterminous United States. *Water Resources Research*, *53*, 7968–7986. https://doi.org/10.1002/2017WR020917

Wing, O. E. J., Bates, P. D., Smith, A. M., Sampson, C. C., Johnson, K. A., Fargione, J., & Morefield, P. (2018). Estimates of present and future flood risk in the conterminous United States. *Environmental Research Letters*, *13*(3), 034023. https://doi.org/10.1088/1748-9326/aaac65

Wing, O. E. J., Pinter, N., Bates, P. D., & Kousky, C. (2020). New insights into US flood vulnerability revealed from flood insurance big data. *Nature Communications*, *11*(1), 1444. https://doi.org/10.1038/s41467-020-15264-2

Wing, O. E. J., Sampson, C. C., Bates, P. D., Quinn, N., Smith, A. M., & Neal, J. C. (2019). A flood inundation forecast of Hurricane Harvey using a continental-scale 2D hydrodynamic model. *Journal of Hydrology X*, *4*, 100039. https://doi.org/10.1016/j.hydroa.2019.100039

Winsemius, H. C., Aerts, J. C. J. H., van Beek, L. P. H., Bierkens, M. F. P., Bouwman, A., Jongman, B., et al. (2016). Global drivers of future river flood risk. *Nature Climate Change*, *6*(4), 381–385. https://doi.org/10.1038/nclimate2893

Winsemius, H. C., Van Beek, L. P. H., Jongman, B., Ward, P. J., & Bouwman, A. (2013). A framework for global river flood risk assessments. *Hydrology and Earth System Sciences*, *17*(5), 1871–1892. https://doi.org/10.5194/hess-17-1871-2013

Winter, T. C. (2001). The concept of hydrologic landscapes. *Journal of the American Water Resources Association*, *37*(2), 335–349. https://doi.org/10.1111/j.1752-1688.2001.tb00973.x

Wolock, D. M. (2003) Hydrologic landscape regions of the United States. *US Geological Survey Open-File Report* 2003-145. https://doi.org/10.3133/ofr03145

Wu, H., Adlter, R. F., Tian, Y., Huffman, G. J., Li, H., & Wang, J. (2014). Real-time global flood estimation using satellite-based precipitation and a coupled land surface and routing model. *Water Resources Research*, *50*, 2693–2717. https://doi.org/10.1002/2013WR014710

Wyncoll, D., & Gouldby, B. (2015). Integrating a multivariate extreme value method within a system flood risk analysis model. *Journal of Flood Risk Management*, *8*(2), 145–160. https://doi.org/10.1111/jfr3.12069

Yamazaki, D., de Almeida, G. A. M., & Bates, P. D. (2013). Improving computational efficiency in global river models by implementing the local inertial flow equation and a vector-based river network map. *Water Resources Research*, *49*, 7221–7235. https://doi.org/10.1002/wrcr.20552

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., & Pavelsky, T. M. (2019). MERIT Hydro: A high-resolution global hydrography map based on latest topography datasets. *Water Resources Research*, *55*, 5053–5073. https://doi.org/10.1029/2019WR024873

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., et al. (2017). A high-accuracy map of global terrain elevations. *Geophysical Research Letters*, *44*, 5844–5853. https://doi.org/10.1002/2017GL072874

Yamazaki, D., Kanae, S., Kim, H., & Oki, T. (2011). A physically based description of floodplain inundation dynamics in a global river routing model. *Water Resources Research*, *47*, W04501. https://doi.org/10.1029/2010WR009726

Yamazaki, D., Lee, H., Alsdorf, D. E., Dutra, E., Kim, H., Kanae, S., & Oki, T. (2012). Analysis of the water level dynamics simulated by a global river model: A case study in the Amazon River. *Water Resources Research*, *48*, W09508. https://doi.org/10.1029/2012WR011869

Yamazaki, D., Sato, T., Kanae, S., Hirabayashi, Y., & Bates, P. D. (2014). Regional flood dynamics in a bifurcating mega delta simulated in a global river model. *Geophysical Research Letters*, *41*, 3127–3135. https://doi.org/10.1002/2014GL059744

Zischg, A. P., Felder, G., Weingartner, R., Quinn, N., Coxon, G., Neal, J., et al. (2018). Effects of variability in probable maximum precipitation patterns on flood losses. *Hydrology and Earth System Sciences*, *22*(5), 2759–2773. https://doi.org/10.5194/hess-22-2759-2018